



PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

**A SLIDING-BOX APPROACH TO DETECTING
PEOPLE IN IMAGES OF INDOOR ENVIRONMENTS
USING WIDE-BASELINE STEREO CAMERA
SYSTEMS**

CHRISTIAN PHILIP PIERINGER BAEZA

Thesis submitted to the Office of Graduate Studies
in partial fulfillment of the requirements for the Degree of
Doctor in Engineering Science

Advisor:
DOMINGO MERY

Santiago de Chile, January, 2015

© MMXV, CHRISTIAN PHILIP PIERINGER BAEZA



PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE
SCHOOL OF ENGINEERING

**A SLIDING-BOX APPROACH TO DETECTING
PEOPLE IN IMAGES OF INDOOR ENVIRONMENTS
USING WIDE-BASELINE STEREO CAMERA
SYSTEMS**

CHRISTIAN PHILIP PIERINGER BAEZA

Members of the Committee:

DOMINGO MERY

ALVARO SOTO

MIGUEL TORRES

JAVIER RUIZ

RENÉ VIDAL

CRISTIAN VIAL

Thesis submitted to the Office of Graduate Studies
in partial fulfillment of the requirements for the Degree of
Doctor in Engineering Science

Santiago de Chile, January, 2015

*In memory of my father, my best friend,
who taught me to lift the head up, and
never lose the faith ...*

ACKNOWLEDGMENTS

Throughout this journey, there are ties that go beyond that the process of growth as a doctoral student. What is visible and permanent are those persons with whom I shared a friendship, long hours of work, successes and failures.

I want to thank my advisor Prof. Domingo Mery for relying on my work and for his tireless support during my research. He provided me with the main tools for carrying out my research and for my future challenges. Thanks to Alvaro Soto for his help in the development of the models used in this thesis and the improvements of my research skills. I also want to thank the members of the committee for supporting this proposal, and for their reviews which help me to improve this document.

To my doctoral friends Billy, Vladimir, Miguel, Tomás, Hans, Pablo, Karim, and Alberto with whom I shared my work, subjects and courses. I thank you for all your friendship, support, and for showing me that when you cannot win at everything you can still laugh at everything.

My family was a special support for me during this process. I thank to my mother Angélica, and my brothers Johnny and Oskar, for their prayers, support and love at distance. To my father, Jonny, who would have loved to see this work completed. I thank you for all you done and all we share. To my son, Dante, thank you to keep my feet on the ground and teach me everyday the simple things of the life. To Laura, thank you to show me the world with a smile. And I specially thank so much to my wife, Isabel, for her unconditional support and love. Thank you for trusting in my decision and accompanying me in this adventure.

Contents

ACKNOWLEDGMENTS	iv
List of Figures	viii
List of Tables	xv
RESUMEN	xvi
ABSTRACT	xviii
Chapter 1. INTRODUCTION	1
1.1. Hypothesis	4
1.2. Objectives	5
1.2.1. General Objective	6
1.2.2. Specific Objectives	6
1.3. Summary of Contributions	7
1.4. Document Organization	7
Chapter 2. BACKGROUND	9
2.1. Image Features	9
2.1.1. Sparse Local Sampling	10
2.1.2. Dense Sampling	12
2.2. Classification Methods	15
2.2.1. Support Vectors Machines	16
2.2.2. Boosting Classifiers	16
2.3. People Detection	17
2.3.1. Single View Approaches	17
2.3.2. Multiple View Approaches	19
2.3.3. Related Approaches for Detection	21
2.4. Non-Maximal Suppression	25
2.5. Focus-of-Attention	26
2.6. Discussion	30

Chapter 3. PROPOSED APPROACH	33
3.1. Spatial Focus-of-attention	36
3.2. Multi-view Projection	37
3.3. Feature Extraction	41
3.4. Multiple View Classifier	44
3.4.1. Ensemble of Features	45
3.4.2. Ensemble of Classifiers	46
3.4.3. Bootstrap Training	48
3.4.4. Best Collection of Aspects	49
3.5. Non Maximal Suppression	50
Chapter 4. METHODOLOGY AND IMPLEMENTATION	54
4.1. Hardware	54
4.2. Datasets Details	54
4.2.1. Train Dataset	55
4.2.2. Test Dataset	58
4.3. Evaluation Methodology	59
4.3.1. Detection Evaluation	60
4.3.2. Heat Maps	61
4.3.3. Per-Windows Evaluation	61
4.4. Publications	61
4.4.1. Flaws Detection	62
4.4.2. Head Modeling	63
4.4.3. Head Detection Using Sliding-Boxes In Multiple Views	64
Chapter 5. EXPERIMENTS AND RESULTS	66
5.1. Implementation Details	66
5.2. Detection Performance	67
5.3. Enriched features	79
5.4. Focus-of-Attention	79
5.5. Using of Geometrical Cues to Detection	81
Chapter 6. DISCUSSION	85

6.1. Conclusions	85
6.2. Future Work	89
References	92
APPENDIX A. Flaw Detections In Aluminium Die Casting	112
APPENDIX B. Head Modeling	126
APPENDIX C. Head Detection Using Sliding-boxes in Multiple Views	137

List of Figures

1.1	General framework and flow of visual surveillance based on image processing. This process involves N cameras and six optional steps for each device. Approaches based on background subtraction involve steps related to environmental modeling and motion segmentation. The fusion of information follows from any of these six steps.	2
1.2	Diagram of the proposed approach for detecting people in a multiple view configuration. Our approach requires N calibrated cameras C_1, \dots, C_N . In this example the sliding-box \mathbf{B} runs through various positions (X, Y, Z) , scanning the entire space S of the scene in which heads could be located. The sliding-box B is projected from 3D space onto images I_1, \dots, I_N , retrieving N detection windows W_{i1}, \dots, W_{iN} . Using 3D information allows us to directly scan the image at a suitable scale for human heads.	6
2.1	Examples of image sampling methods. (a) sparse local sampling: image regions are selected using a salient region detector. (b) dense representation: the image is sampled using a dense grid.	10
2.2	Examples of methods for sampling images. (a) Multi-scale pyramid decomposition sampling strategy: the image is downsampled and smoothed at every level and then sampled at smaller subregions. (b) Spatial pyramidal decomposition strategy: the image is sampled at smaller subregions by dividing the image until a maximum number of levels is achieved.	15
2.3	Examples of an aspect-graph. (a) Two views or aspects of a polyhedral object. These aspects correspond to projected faces, which form a graph according with their adjacency, <i>i.e.</i> , faces 1, 2, and 4 in the Aspect 1 appear connected because they are adjacent. (b) The viewing sphere sampled at regular intervals. During sampling, images are captured every five degrees. An iterative procedure combines views into aspects using prototypes which represent each aspect.	23
3.1	(a) Example of the aspects collected from multiple view projections using our proposed approach 1. The set of projections form a collection of aspects of the head. We use this projection strategy to collect all of the information available from all of the viewpoints in	

the camera system. (b) Block diagram of the proposed method. Our approach includes five main steps: spatial focus-of-attention, multi-view projection, feature extraction, classification and non-maximal suppression. In the example, we use $N = 4$ cameras. The algorithm begins with an input scene composed of I_1, \dots, I_4 images. Then, the spatial focus-of-attention reduce the number of head hypotheses. Next, the algorithm computes the projections of the box \mathbf{B}_i onto the image I_j as W_{ij} , forming a collection of aspects. Afterwards, we extract a set of features for each projection W_{ij} and apply the model for sequences. Finally, a spatial NMS procedure allows us to eliminate multiple detections. 35

3.2 Explanation of the focus-of-attention procedure. (a) Blue dots show the potential head positions detected by our focus-of-attention procedure. This process provides hypotheses with more likely location of heads within the region of interest S and helps us to drastically filter the spatial detections. (b) Yellow circles shows ground-truth heads examples within the subspace S 37

3.3 Shows the triangulation between head hypotheses from two images I_1 and I_2 at different viewpoints. Hypotheses $\{\mathbf{h}_{11}, \mathbf{h}_{21}, \mathbf{h}_{31}\}$ in image I_1 generate epipolar lines l_1, l_2 and l_3 in image I_2 . The pairs set of head hypotheses $\{\mathbf{h}_{11}, \mathbf{h}_{12}\}$ and $\{\mathbf{h}_{21}, \mathbf{h}_{22}\}$ share the same 3D position $\hat{\mathbf{M}}_1$ and $\hat{\mathbf{M}}_2$, respectively. We estimate these spatial positions by triangulation using least square minimization along the ray $\overline{h_{ij}C_i}$. Due to there are no head hypothesis near to epipolar line l_3 , the \mathbf{h}_{31} do not generate potential head position. 38

3.4 Projection diagram of a sphere quadric \mathbf{Q} defined on \mathbf{M}_i with radius r . In this example, for $N = 4$, \mathbf{Q} is projected onto the images I_1, \dots, I_4 as a conic \mathbf{C} . The projections W_{ij} are defined as the maximum quadrilateral subscribed over \mathbf{C} and showed as dashed red circles in this figure. The elements W_{ij} represent the projection of \mathbf{B}_i onto the camera j . All of these elements define a collection of aspects which represents the box \mathbf{B}_i seen from each camera. Each element W_{ij} was cropped and then rescaled to 64×64 pixels before feature extraction to cope with projection at different size. 41

3.5 Diagram of the space of interest S inside of a room and defined as parallelepiped with set of boundaries $[X_a, X_b]; [Y_a, Y_b]; [Z_a, Z_b]$. We use this contextual information to limit the action of our sliding-box \mathbf{B}_i within the space S . This allows to us to search for people's heads in areas in which they are likely to appear according to the context. 42

3.6	Example of the pyramidal feature extraction on a W_{ij} patch. Features are computed in $l = 0, 1, 2$ levels of the image patch. In each level, the image has 4^l cells. The final descriptor has $N_f = 59 \times (1 + 4 + 16) = 1,239$ bins.	43
3.7	Example of the pyramidal feature extraction on a W_{ij} patch. Features are computed in $l = 0, 1, 2$ levels of the image patch. In each level, the image has 4^l cells. The final descriptor has $N_f = 59 \times (1 + 4 + 16) = 1,239$ bins.	44
3.8	Training process diagram of features ensemble scenario. Once we extract features from each element W_{ij} , we concatenate all of the N_f features in a single descriptor with N_s bins. We use a single SVM classifier to learn a model β_{fe} using examples of head sequences.	47
3.9	Training process diagram for the classifier ensemble scheme. We use an ensemble of classifiers divided into two layers: (a) shows the first layer, which is formed by multi-class SVM classifiers. This layer can identify frontal head, rear head, and background. (b) shows the second layer, which is trained using the scores $(f_{\beta}^1, \dots, f_{\beta}^k)$ obtained by applying the first layer of classifiers to each element W_{ij} . This process yielded the model β , which can classify the image sequence.	49
3.10	Diagram of best collection searching. The algorithm receives an input collection of aspects without a priori knowledge about its correct alignment. We apply a set of circular shifts to generate the total number of four collections of aspects, including the input collection. After applying the multiple view classifier to each collection, we choose the most confident one using an <i>argmax</i> criterion. In the diagram, the multiple view classifier assigns a set of confidence values to each collection of aspects: $f_{\beta}(x_i^0) = -1.25$, $f_{\beta}(x_i^{+1}) = -0.12$, $f_{\beta}(x_i^{+2}) = 1.75$, $f_{\beta}(x_i^0) = 0.75$. Finally, our algorithm selects the collection of aspects generated with the second shift because it best matches the training samples.	51
4.1	We mounted four Point Grey Flea2 cameras in a classroom to simulate an indoor environment where we could detect people. The four cameras were calibrated to ensure the geometric reconstruction of 3D points and synchronized to acquire the images at same time.	55

4.2	Example of our GUI for labeling a set of four images representing the same scene at different viewpoints. A head is labeled selecting two matching points and we estimate its 3D location throughout the geometric model. This location is re-projected onto the images as bounding boxes which represent the head in all viewpoints.	56
4.3	Example of images collected for training. Each image comes from one view in the camera system. People stand over the white crosses on the floor and spin around their Z axis to generate various views of the head.	57
4.4	Examples of aspects collection used for training. (a) positive instance of people's head retrieved from multi-view camera system and (b) background samples of classroom environment. We used four cameras in both examples, where $j = 1, \dots, 4$	58
4.5	Example of images collected for testing. Both test datasets contain images of people in a classroom at different activity levels and under various occlusion conditions. (a) The first test sequence contains people moving and changing their appearances. (b) The second test sequence consists of people sitting observing a lecture. Their appearances change less than the first sequence, and although there is occlusion, it is less frequent.	59
4.6	Example of heat map used to analyze the most frequently visited areas. This map shows a top-view of the classroom used for our experiments. Pseudo color indicates the number of detections at each location. Red areas are equivalent to a high number of detections, blue areas point a low number of detections. Similar as we describe in Fig. 3.5, the space of interest S is limited by the yellow square. The dashed region marks the space that we do not take into account during detection.	62
4.7	Concept testing for our approach, applied to flaws detection in aluminum die casting. We build a flaw sequence using 72 images in a calibrated multiple view system. The classifier learns from simulated flaw sequences and identifies flaws on real images.	64
5.1	Summary of the head model that the DPM learns after training. (a) Example of the model, its gradient map, and parts. (b) Example of detections after applying the model over test images. .	68
5.2	Precision-Recall curves comparison of detection performance in sequence <i>sq-01</i> . We compare performances of our methods EF, OVA-EC, and OVO-EC versus performance of single view DPM and the epipolar version of DPM. In general we report high performance level of our OVO-EC approach. The overall performance is represented using the AP	

	value. (a) Performance evaluation using LMS-NMS. Although detectors performed well, the DPM performed best overall. Our approach based on a one-vs-one ensemble of classifier presented the best performance. We note an improvement in precision on the last part of the curve that allowed to our method to maintain its level of precision at the same recall rates as DPM. (b) Performance evaluation using NMS based in mean-shift procedure. In this case, our approach based on a OVO-EC of classifier performed best. We observed the same improvement in precision on the last part of the curve, but in this case it was more noticeable. In both curves, DPM using epipolar geometry showed marginal improvement after recall of 0.85. This may be due to the way that epipolar geometry helps to eliminate false detections.	71
5.3	Precision-Recall curves comparison of detection performance in dataset <i>sq-02</i> . Comparisons include our three methods EF, OVA-EC, and OVO-EC versus performance of single view DPM detector and the epipolar version of DPM. In general we report best performance level in the OVO-EC approach. The overall performance is represented using the AP value. (a) Performance evaluation using NMS based on greedy procedure. The three multiple view detectors performed at similar similar levels. All of them outperform the DPM detector in its two versions. Our approach based on a OVO-EC yielded the best performance as in <i>sq-01</i> . We note improvements in precision on the last part of the curve which allowed our method to maintain a level of precision at the same recall rates as DPM. (b) Performance evaluation using NMS based on mean-shift procedure. Our approach based on a one-vs-one ensemble of classifiers performed best. The same improvements in precision are present. In both curves, DPM detector using epipolar geometry does not show an improvement. This may be due to the fact detections near to same epipolar lines are eliminated.	72
5.4	Heat-map of detections using our three detection approaches in sequence <i>sq-01</i> after applying WMS-NMS and LMS-NMS. Left column are detection using WMS-LMS and right column are detection using LMS-NMS. (a) Heat-map of the ground-truth. (b) and (c) Heat-maps of detections using the EF approach. (d) and (e) Heat-maps of detection using OVA-EC. (f) and (g) Heat-maps of detection using OVO-EC.	75
5.5	Heat-map of detections using our three detection approaches in sequence <i>sq-02</i> after applying WMS-NMS and LMS-NMS. Left column are detection using WMS-LMS and	

	right column are detection using LMS-NMS. (a) Heat-map of the ground-truth. (b) and (c) Heat-maps of detections using the EF approach. (d) and (e) Heat-maps of detection using OVA-EC. (f) and (g) Heat-maps of detection using OVO-EC.	76
5.6	Detections comparison using WMS-NMS. (a) Detections provided by DPM detector. (b) Detections generated by OVA-EC approach. (c) Detections yielded by OVO-EC approach. (d) Detections generated by applying the EF method. We display detections in red boxes and ground-truth heads in green boxes. The results show that our methods can retrieving heads that the 2D detector missed. Although we retrieve new heads, we also add some noisy detections.	77
5.7	Detections examples using NMS based on mean-shift. (a) Detections provided by Latent-SVM detector. (b) Detections generated using OVA-CE. (c) Detections yielded by OVO-CE. (d) Detections yield by applying FE method. Detections appear as red boxes and ground-truth heads as green boxes. The results show most of the heads missed by the 2D detector were retrieved by our multiple view detector. Although we retrieve new heads we also add some noisy detections due to hallucinations of the classifier.	78
5.8	Performance comparison in per-windows classification. The curves show performance evolution as we add information from the four visual sources of our camera system. (a) precision-recall curve of ensemble of features. This method presents an improvement in performances after the third camera is included, which is the view that includes information about the rear of the head. We conclude that the additional viewpoint contributes to improving the performance. (b) The precision-recall curve of ensemble of classifiers using a one-vs-all strategy. (c) The precision-recall curve of ensemble of classifiers using a one-vs-one strategy. These second method presents improved performance as we add a new viewpoint. It is smoother that the improvement presented by ensemble of features. Overall performance is represented using the AP value. High AP values are yield by all methods when we use the four cameras.	80
5.9	Sensitivity analysis of confidence vs recall due to use of the spatial focus-of-attention procedure. There is a trade-off between detector sensitivity and the maximum recall reached upstream of the classifier. If we set a low sensitivity for the detector, we increase the ability of the focus-of-attention to retrieve all detections in the scene. Even though a	

low confidence threshold produces more spatial hypotheses and increases the burden for other procedures, it is always better use the focus-of-attention than to run the sliding-box across the entire region of interest S . (a) shows sensitivity for sequence sq -01 where the maximum recall appears to fall between 0 and -0.4 of detector confidence. (b) shows sensitivity for sequence sq -02 where the maximum recall appears to be between -0.4 and -0.7 of detector confidence.	82
5.10 Examples of focus-of-attention computed within the region of interest S using the outputs provided by an interest-point detector. The head hypotheses appear as grey circles and the ground-truth as green stars. In accordance with our experiments, the spatial focus-of-attention recover most of the ground-truth elements. Though it adds an overhead, it is always better use the focus-of-attention than to run the sliding-box across the region S . Left column shows focus-of-attention build setting a low threshold in the inters-point detector. Right the focus-of-attention using a higher threshold.	83
5.11 Random sample of 60 windows evaluated: (a) by a 2D detector, and (b) by our approach. In (a), windows must change their size in order to predict the real size of the object. We improve the search by using the real size of the object what limit to the size and locations of the projected windows, as shown in (b).	84
5.12 Alignment test of collection of aspects. The figure shows the results of using sliding-box to evaluate among a set of hypotheses and their scores. Box scores are high when the box belongs to the head class, and its projections yield better alignment, as shown in (a) and (b). In (c) we observe background examples and their scores, all of which were negatives.	84

List of Tables

4.1	Details about the training dataset used to train the feature ensemble. Each instance is a collection of aspects, as shown Fig. 4.4	57
4.2	Details about the training dataset used for training individual models.	58
4.3	Details about test dataset used for testing the detector. Sequences were called sq-01 and sq-02. We show an average number of people per image and the total number of frames in each sequence.	59
5.1	Repeatability analysis of our three approaches in dataset <i>sq-01</i> . Results show little variation after detection. All methods show significance over the 95% of confidence level.	73
5.2	Repeatability analysis of our three approaches in dataset <i>sq-02</i> . The results show little variation in detection that passes the 95% of confidence level.	73
5.3	Summary of the burden reduction due to the spatial focus-of-attention. Although there is an overhead due to applying an interest-point detector, it is always better use this procedure than perform an exhaustive search across the region of interest.	81

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

UN ENFOQUE DE VOLÚMENES DESLIZANTES PARA LA DETECCIÓN DE PERSONAS
EN IMÁGENES DE AMBIENTES INTERIORES USANDO UN SISTEMA MULTICÁMARAS

Tesis enviada a la Dirección de Postgrado en cumplimiento parcial de los requisitos para el grado
de Doctor en Ciencias de la Ingeniería.

CHRISTIAN PHILIP PIERINGER BAEZA

RESUMEN

En las últimas dos décadas ha aumentado masivamente el uso de cámaras en sistemas de vigilancia y monitoreo de actividades, haciendo difícil su seguimiento el 100% del tiempo por operadores humanos. La detección de personas ha provocado gran interés en investigadores de la comunidad de visión por computador, con el fin de generar herramientas de vigilancia automática. Los primeros trabajos de detección se basaron fuertemente en técnicas de procesamiento de imágenes, las que a pesar de su rapidez y simplicidad son sensibles a los cambios de iluminación, oclusión, y variación de las poses humanas. Actualmente, los enfoques de aprendizaje de máquina basados en ventanas deslizantes han tenido éxito significativo en la detección de personas. Este éxito se debe en parte al uso de poderosos modelos de aprendizaje de máquina, características visuales nuevas y más informativas y modelos basados en partes capaces de manejar la variabilidad de los objetos. Un denominador común de estas técnicas es que ellos confían principalmente en métodos de aprendizaje estadístico que usa información de la intensidad de las imágenes para capturar las características de apariencia de los objetos. Una limitación importante de estos enfoques basados en apariencia es que no incorporan información geométrica relevante que provea pistas espaciales tales como el tamaño real de los objetos a detectar, profundidad o la ubicación más probable de estos objetos en la escena. Algunos trabajos recientes consideran el beneficio de incorporar información de varios puntos de vista. La detección usando una sola cámara es apropiada cuando existe oclusión leve, sin embargo, para casos de mayor oclusión el uso de múltiples vistas permite mejorar la detección.

A pesar de que existen técnicas para relacionar la información en múltiples vistas, aún quedan desafíos importantes que resolver. En esta tesis, proponemos un enfoque para detección de personas que une avances en detección basada en aprendizaje de máquina con geometría de múltiples vistas. La idea principal de nuestro método es barrer un volumen virtual a través del espacio con el fin de analizar solo la parte de las imágenes donde este elemento es proyectado. Este esquema nos permite resolver problemas relacionados al establecimiento de correspondencias entre cámaras, incluir información espacial, y enriquecer los modelos de detección usando características enriquecidas. Este documento describe nuestro enfoque y su evaluación en detección de personas en ambientes interiores. Los experimentos demuestran que nuestro método mejora detectores 2D del estado del arte en 10% respecto del *precision-recall* promedio de su mejor vista, usando iguales condiciones de entrenamiento. Los resultados muestran que nuestro enfoque puede ser usado efectivamente para detección de personas en sistemas de múltiples vistas.

Palabras Claves: detección de personas, ventanas deslizantes, ambientes interiores, múltiples vistas, red de cámaras, SVM.

Miembros de la Comisión de Tesis Doctoral

DOMINGO MERY

ALVARO SOTO

MIGUEL TORRES

JAVIER RUIZ

RENÉ VIDAL

CRISTIAN VIAL

Santiago, January, 2015

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
COLLEGE OF ENGINEERING

A SLIDING-BOX APPROACH TO DETECTING PEOPLE IN IMAGES OF INDOOR
ENVIRONMENTS USING WIDE-BASELINE STEREO CAMERA SYSTEMS

Thesis submitted to the Office of Graduate Studies in partial fulfillment of the requirements for the
Degree of Doctor in Engineering Sciences by

CHRISTIAN PHILIP PIERINGER BAEZA

ABSTRACT

Over the past two decades, there has been a massive increase in the use of digital cameras in surveillance systems and monitoring activities. This has made it difficult for human operators to provide 100% coverage at all times. The ability to detect human forms in video images has generated a great deal of interest among researchers in the computer vision community who are working on the design of automatic visual surveillance tools. Previous studies of this topic were strongly based on image processing techniques, which in spite of their speed and simplicity are sensitive to changes in lighting, occlusion and the variability of human poses. Recently, machine learning approaches based on sliding windows have proven to be successful in people detection. This significant success is due in part to the use of powerful machine learning models, new and more informative visual features, and part-based models which cope with object variability. A common denominator of these techniques is that they rely mainly on statistical learning methods that exploit image-intensity information to capture object appearance features. An important limitation of appearance-based approaches is that they do not incorporate relevant geometric information that can provide important, useful spatial cues such as the real size of the object to be detected, depth, and likely spatial appearance location in the scene. A few recent works consider the benefits of including information from various viewpoints. Detection using one camera is suitable when there is mild occlusion, however, if there is heavy occlusion multiple views help to improve final detection. Despite the existence of techniques for linking information across multiple views, significant challenges remain. In this thesis, we propose a multiple view detection approach in order to bridge

the gap between advances in machine learning-based object detection and multiple view geometry. The key idea is to run a virtual volume across the space in order to analyze only the corresponding portion of the images where this 3D element is projected. This allows us to solve problems related to correspondence among cameras. We also can include useful spatial cues and enhance detection models with enriched features descriptors. This document describes our approach to people detection in video images of indoor environments and its evaluation. The experiments show that our framework improves detection levels of 2D state-of-the-art methods in 10% of the average precision-recall at their optimal view using the same training conditions. These results suggest that our approach can be used effectively to detect objects in multiple views.

Chapter 1. INTRODUCTION

Over the past two decades, there has been an increase in the number of digital video cameras used in security and monitoring systems. The early devices were only used in private surveillance and retail environments. Over time, their applications shifted towards the area of public space surveillance (Kuno et al., 1996; Dee & Velastin, 2007). Initially, security personnel could monitor these devices easily due to the small number of cameras used. However, the intensive use of these devices has made it impractical to depend on human operators to provide all monitoring services all of the time. Two main factors limit the practicality of the exclusive use of human monitoring: the amount of information gathered and limited attention spans and concentration. The use of techniques from the fields of image processing, computer vision and machine learning has allowed for the design of intelligent systems which help operators engage in automated monitoring using visual information. According to Hu et al. (2004); Valera and Velastin (2005); Cristani et al. (2010), the basic elements of an automated vision system are the environment model and motion segmentation, object detection or object classification, tracking, behavior interpretation and personal identification, and fusion of information from multiple cameras (see Fig. 1.1). The major tasks developed for these systems are people and object detection, detection of abandoned objects, and the tracking of specific subjects (Hu et al., 2004). The goal is to generate complex information about scenes such as the number of people present, the identities of those individuals, activity and behavior recognition, borders of hazardous or forbidden areas, and flow analysis. One important aspect of this field of inquiry is the evaluation of these systems. Influential conferences such as PETS, CREDS, i-LIDS, ETISEO, and PASCAL have set forth metrics which allow researchers to compare their algorithms performance (Dee & Velastin, 2007).

Early work on detection was strongly based on image processing techniques. The main task at that time was to separate foreground objects from the background. The essential stage of the identification of these elements is the background subtraction process. Once this segmentation is complete, the objects appear as blobs that can be classified according to shape, color, movement or another distinctive feature. Because the segmentation process uses consecutive frames, this approach is commonly used in video sequences. The most frequently cited studies for background subtraction address frame differences, Gaussian moving average Wren et al. (1997), mixture of

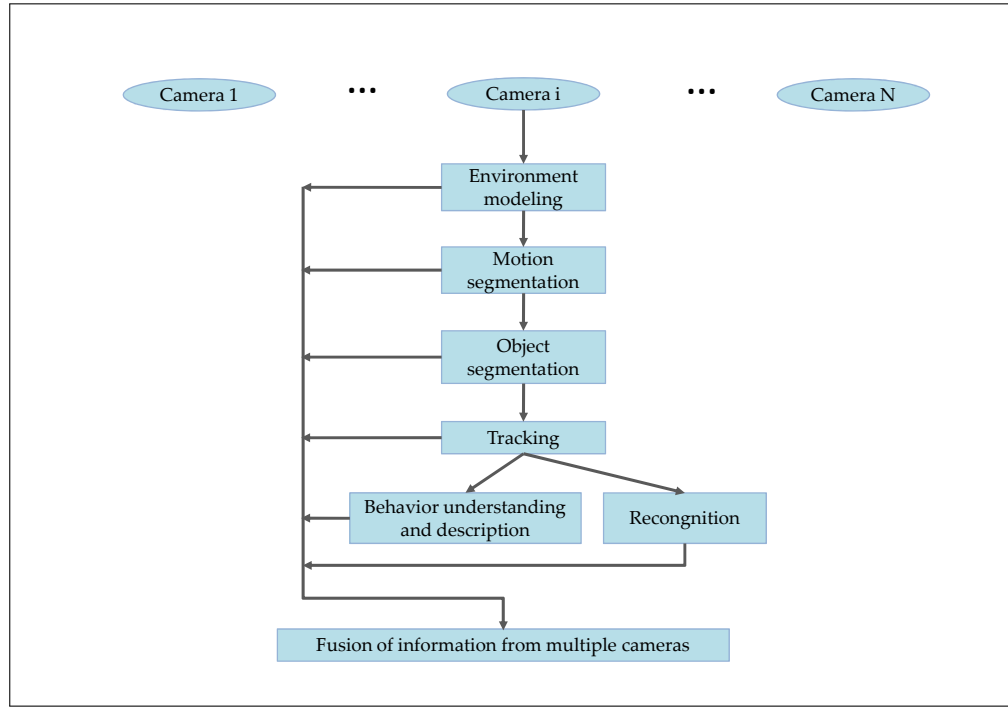


Figure 1.1. General framework and flow of visual surveillance based on image processing. This process involves N cameras and six optional steps for each device. Approaches based on background subtraction involve steps related to environmental modeling and motion segmentation. The fusion of information follows from any of these six steps.

Gaussians Stauffer and Grimson (1999), temporal median filter Lo and Velastin (2001), and code-book models Kim et al. (2005). Background subtraction is a simple and computationally efficient method, but it requires background initialization and background updating. Its main challenges and drawbacks are related to objects that have moved, time of day, light changes, oscillating objects, occlusions or camouflage, bootstrapping, foreground aperture, sleeping foreground or persistent objects, shadows, and reflections (Cristani et al., 2010).

A new generation of algorithms based on machine learning techniques is addressing the task of detecting people in static or still images. These methods have received a great deal of attention due to their performance level and the independence of the segmentation processes, which is sensitive to lighting changes. This significant progress in people detection is due in part to the use of powerful machine learning models, new more informative visual features, and part-based models which cope with the objects variability (Viola et al., 2005; Dalal & Triggs, 2005; Felzenszwalb et al., 2009; Girshick et al., 2014; Dean et al., 2013). In general, these algorithms use a sliding-window approach

to sample the input image at several scales. Then every window is represented by a set of features or measurements coded as a vector that describes shape, color, texture, etc. Afterwards, windows are classified according to an object target class using a previously learned model. Finally, a non-maximal suppression step avoids the multiple detections. The most successful detection algorithms use local descriptors in order to extract global and local spatial information from each instance and to cope better with inter-class variability (Ojala et al., 2000; Lowe, 2004; Mikolajczyk & Schmid, 2004; Dalal & Triggs, 2005; Lazebnik et al., 2006; Bosch et al., 2007). Though most of this algorithms have a high computational demand during training, some allow for real-time detection and can be used through frames in video sequences. As stated in Dollar et al. (2011), although there have been some detection improvements the overall performance is still poor.

A common denominator of those techniques is that they mainly rely on statistical learning methods that exploit image-intensity information to capture object appearance features. Their goal is to uncover visual spaces where visual similarities carry enough information to achieve robust visual recognition. As a relevant limitation, appearance-based approaches do not incorporate relevant geometric information that can provide useful and relevant spatial cues such as the real size of the object to be detected, depth, and spatial likely appearance location. There are some notable exceptions of approaches that combine detection based on machine learning algorithm with additional spatial information about the objects in order to discard false detections (Helmer & Lowe, 2010; Salas & Tomasi, 2011; Spinello & Arras, 2011; Espinace et al., 2013). Results show that there have been improvements in regard to recovering the spatial information lost during image acquisition. Though they do have some advantages, these methods are focused on mobile robots and require supplementary hardware such as stereo and depth cameras to recover these cues. The large number of cameras installed today makes it possible to obtain information from different viewpoints simultaneously and to exploit the same spatial cues, establishing relationships across the views in a camera system. This provides opportunities to combine and integrate information from various visual sources.

Detection using one camera is suitable when there is mild occlusion, but in situation of heavy occlusion the relationships between multiple views contributes to identify targets across views when the visual appearance alone is insufficient (Mittal & Davis, 2003; Khan & Shah, 2006; Eshel & Moses, 2010; Liem & Gavrila, 2013; M. Song et al., 2013). Some researchers have proposed

using geometric techniques to establish relationships across views in a camera network, providing efficient ways to combine and integrate this information (Szeliski, 2010). Detection in multiple camera configuration present two main challenges. On the one hand, there are practical constraints related to the establishment of correspondence between observations when the appearance of the target object change drastically among viewpoints. And on the other hand, difficulties arise in non-overlapped configurations when similar appearances are simultaneously present in the camera network (M. Song et al., 2013).

Although there are well-established detection approaches based on machine learning techniques, the current methods still present several drawbacks that must be overcome. On the one hand, we observe that in general single view approaches to detecting people or objects mainly *i)* use a sliding-window at various scales to compensate scale changes of the object target class in images that produces false positives due to hallucinations at several scales; *ii)* do not take into account the use of additional cameras to improve the overall detection; and *iii)* do not take into account useful 3D information such as real sizes of people or objects, and the positions in which the target object classes are likely to be found at the scene. On the other hand, wide baseline stereo systems present limitations related to correspondence matching in cases in which the same object has various poses or variations simultaneously. We address some of these issues by proposing an approach based on the idea of enhancing the total amount of information available in a camera system, and using the advantages from machine learning and multiple view geometry.

1.1. Hypothesis

Computer vision applications based on multiple views use geometric rules to determine the structure of the scene. This set of rules allow us to establish relationships among the elements in each view of the camera system. It can be also generated using prior knowledge from the vision system itself such as its projection matrices and a calibration model, or they can also be estimated using landmarks from the scene. However in this last case, the metric information cannot not be directly recovered (Hartley & Zisserman, 2003; Szeliski, 2010).

A calibration model provides a geometric relationship between the world coordinates and the images, and relationships among cameras in a multiple views configuration. For example in the detection problem, this geometric structure would allow us to filter out false detections present

in one view when detections are not consistent in the remaining views. Correspondence analysis increases total detection performance through collaboration among all of the cameras in the vision system. In addition, knowledge of the geometric structure of the scene and the objects within it can add useful information related to their actual size and most likely location. Using this knowledge, we can exclude people or objects outside of a specified area and eliminate detections that have been included by mistake when the 2D detector tries to predict the real size of the objects. Unfortunately, if there are significant differences in appearance between one viewpoint and the next, matching algorithms cannot guarantee that correspondence will be established correctly.

We propose a method for detecting people in a calibrated multiple view system in which a volume element is tailored to size as the object target class to be detected. We call this volume a *sliding-box*, and we define it as a 3D virtual volume element \mathbf{B} as shown in Fig. 1.2. This box passes through the three directions (X, Y, Z) of the spatial domain in the scene where people are likely to appear. Our approach is designed to inspect the corresponding portion of the images where this volume has projections according to its size, Fig. 1.2. This method is based on the standard sliding-window approach in which a detection window runs through an input image in the 2D domain, with both horizontal and vertical direction at various scales (Viola & Jones, 2004; Dalal & Triggs, 2005). Instead, we propose a sliding-box in 3D at fixed size. This allow us to avoid searching at various scales and combine information in a multiple views camera system. However, the same idea could be used in a single 2D detector to guide the scale changes. Thus, our hypothesis is as follows:

The use of a sliding-box allows us to generate a set of potential candidates according to the physical dimensions and positions of the people in the image. In doing so, it allows information from various viewpoints to be combined using correspondent regions. These properties make it possible to increase the effectiveness of detection relative to the single view detection described in the state-of-the-art with and without the use of multiple views.

1.2. Objectives

This section presents the general and specific objectives of the study. We will first present the general objective. Specific objectives will be described throughout the study.

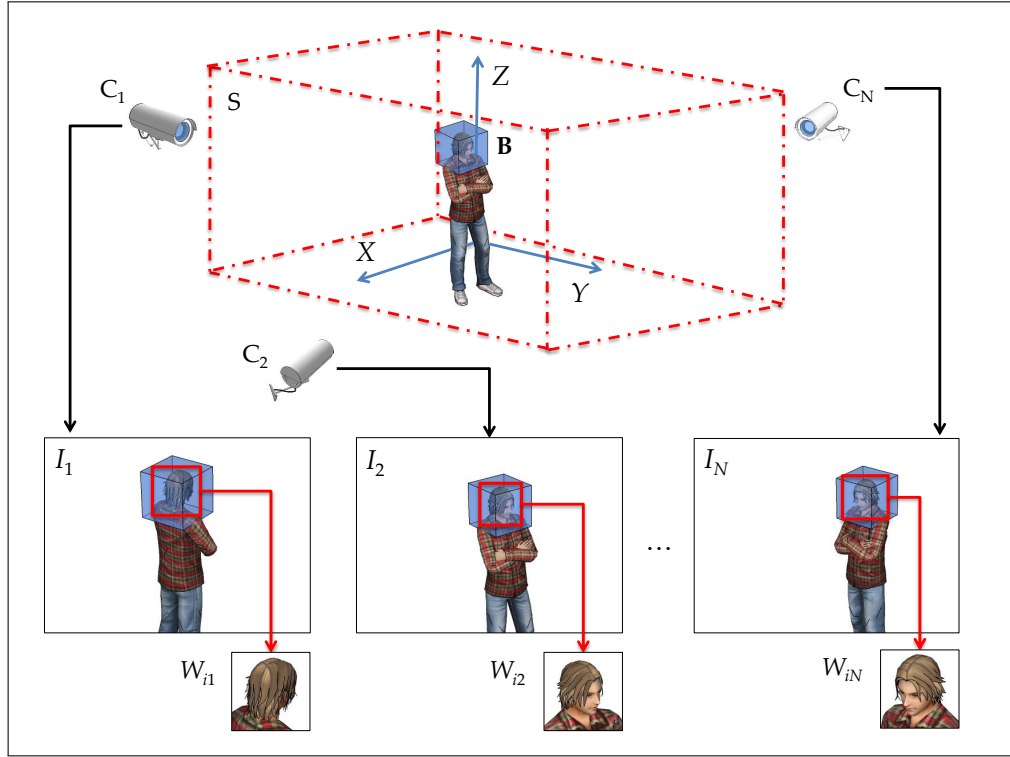


Figure 1.2. Diagram of the proposed approach for detecting people in a multiple view configuration. Our approach requires N calibrated cameras C_1, \dots, C_N . In this example the sliding-box B runs through various positions (X, Y, Z) , scanning the entire space S of the scene in which heads could be located. The sliding-box B is projected from 3D space onto images I_1, \dots, I_N , retrieving N detection windows W_{i1}, \dots, W_{iN} . Using 3D information allows us to directly scan the image at a suitable scale for human heads.

1.2.1. General Objective

According to the hypothesis, the main objective of this thesis is to develop a framework for detecting people in images of indoor environments that allows for the information coming from various viewpoints be combined without depending on correspondence establishing among detections, enriching the information used by a detection algorithm. The results of this framework is the increase of the detection performance.

1.2.2. Specific Objectives

In order to achieve our main objective, we must meet the specific objectives listed below:

- build a multiple view environment for testing in order to imitate indoor conditions,

- build a multiple view dataset for training and testing classifiers according to the environment,
- investigate and design algorithms for people detection in single and multiple views,
- develop an algorithm for detecting people in video images of indoor environments based on the simultaneous projection of a virtual volume which allows us to include spatial and appearance information during the detection process; and,
- extend the proposed approach to other detection problems.

1.3. Summary of Contributions

The proposed approach offers several promising advantages in object detection, including the following three main contributions of this thesis:

- (i) A method that uses geometric information to improve the traditional sliding-window approach applied to current appearance-based detectors, but a sliding box approach that offers the following advantages: *a)* we reduce correspondence problem at level of detection object by projecting the sliding-box onto multiple views simultaneously, instead of using matching methods based on low-level pixel or interest points; *b)* we include useful 3D cues that allow us to focus in the relevant parts of the 3D world in order to filter false candidates; and *c)* we guide the search for people by moving the sliding-box within a 3D space of interest, avoid searching in places where people must not be detected.
- (ii) A classification approach based on combining information from multiple views. This allows us to enrich the data used to train the models. Thus, we are able to include all of the visual information from different visual sources in a single model.
- (iii) Verification of the relevance of the previous ideas for the case of people counting using head detection showing that the proposed approach provides a substantial increase in recognition performance with respect to alternative state-of-the-art techniques.

1.4. Document Organization

This study is divided into six main chapters and three appendixes with publications related to this research. The contents of each chapter are described below:

- Chapter 1 presents the motivations for this work, an overview of our approach, and a summary of the contributions of our work.
- Chapter 2 describes the theoretical foundations needed to develop the process described and includes a review of previous studies related to people detection in single and multiple views.
- Chapter 3 provides a detailed discussion of the proposed method for detecting people using multiple views.
- Chapter 4 describes the methodology used to verify our hypothesis and provides a discussion of its implementation.
- Chapter 5 presents experimental analysis of this work, which was applied to people detection in images of indoor environments.
- Chapter 6 presents conclusions related to the work developed in relation to the specific objectives of this study and discusses areas that may be explored in the future.
- Appendix A presents a concept and its results to apply our approach on flaws classification to aluminum die casting. This text has been accepted for publication as an article in INSIGHT journal.
- Appendix B describes a preliminary approach to modeling human heads. This results has been accepted for publication as an article in the Chilean Workshop of Pattern Recognition (CWPR, 2012).
- Finally, Appendix C presents an summary of the results and conclusions included in this research. This text has been submitted to the journal of Machine Vision and Applications.

Chapter 2. BACKGROUND

Modern visual object detection systems are based on machine learning algorithms that analyze images and generate outputs at locations where the confidence or probability to the target object classes is sufficiently high. These machine learning algorithms generally present four stages: *i)* collection of data from the environment, *ii)* feature extraction, *iii)* selection of the most relevant features to the process, and *iv)* classification. The end result of this process is the detection of the object (Szeliski, 2010; Murphy, 2012).

Features and the strategy used to extract these measures plays an important role in the final performance of the algorithm. Both the strategy and the type of features determine the design of the learning algorithm. Researchers have proposed novel and powerful learning algorithms that can deal successfully with the high dimensional descriptors currently used in object detection. However, detection based on visual information can be improved using geometric information about the scene and the object present in. New approaches include spatial information lost during image acquisition that allows them to eliminate noisy detections. These studies have paved the way for further analysis conducted in an effort to provide the detection system with these spatial cues. We review the main theoretical aspects of this study in the following four sections: image features, classification methods, people detection, and discussion.

2.1. Image Features

Raw images are arrays of pixels that contain light information of a scene picked up by sensor in a digital camera. This large set of data require to be pre-processed in order to transform the visual information into a new space of variables less redundant and more informative than the image domain. The measures used for compressing the visual information are commonly called features. These features need to be discriminative to allow for making the distinction between different classes, while providing invariance to light changes, noise and differences in viewpoint (Szeliski, 2010; Nixon & Aguado, 2012). Advanced detection algorithms based on machine learning use local image descriptors to cope better with objects variation. Further, several kinds of strategies can be used to sample the image and represent these image descriptors. Without loss of generality, we will divide those strategies into two groups based on how they sample the input image: sparse local sampling, dense sampling of image regions. In addition, dense sampling can also be divided

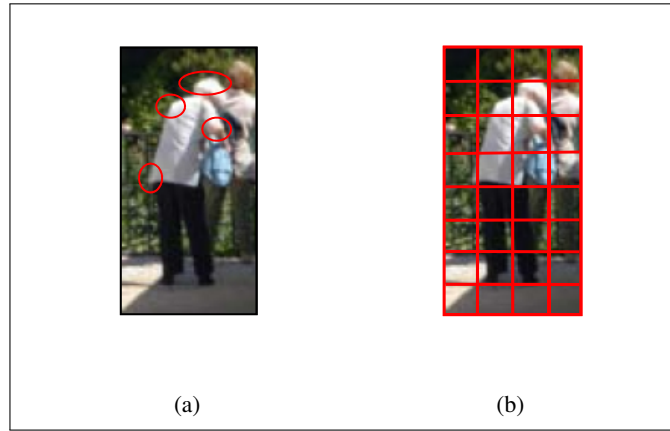


Figure 2.1. Examples of image sampling methods. (a) sparse local sampling: image regions are selected using a salient region detector. (b) dense representation: the image is sampled using a dense grid.

into two categories: sampling using multi-scale pyramid decomposition and sampling using spatial pyramidal decomposition.

reduce the dimensionality and compress the redundant data into a compact representation called features. After the feature extraction,

2.1.1. Sparse Local Sampling

Sparse sampling takes relevant local image regions from the image, as shown in Fig. 2.1a. These salient regions are selected using either key point detectors or parts detectors. The key point detectors select more informative local regions, which are more stable, repeatable and reliable. These factors directly impact the overall detector performance. The most relevant key point detectors are: Difference of Gaussians (DoG) (Lowe, 2004), invariant Harris-Laplace (Mikolajczyk & Schmid, 2004), Maximally Stable Extremal Regions (MSER) (Matas et al., 2004), and affine invariant salient regions (Kadir et al., 2004). In general, these keypoints perform well in problems in terms of establishing matches in order to compose image panoramas. However, saliency detectors work best in object classification (Mikolajczyk & Schmid, 2005; Savarese & Fei-Fei, 2007).

Once the regions have been selected, descriptors are computed over them. The most influential descriptor for sparse representation are the Scale Invariant Feature Transformation (SIFT) (Lowe, 2004) and the Speeded-Up Robust Feature (SURF) (Bay et al., 2008). The SIFT descriptor contains a set orientation histogram weighted on gradient magnitude and the region scale. These gradients are

computed over rectangular grids. The SURF descriptor goes further and improve some features of SIFT descriptor that let it to speed-up the computation performance. This algorithm approximates the DoG using a set of box filters based on a sum of 2D Haar wavelet responses, allowing that the convolution to be computed by summations over integral images as proposed Viola and Jones (2001).

In 2010, Calonder et al. use the same idea of representing salient regions by a descriptor and propose a binary string descriptor called BRIEF. This method is able to directly compute binary strings from image patches avoiding to calculate the full descriptor before further processing. It also includes a Gaussian smoothing that reduce the effect of noise sensitivity in complex scenes. The algorithm computes the binary string descriptor via the intensity comparison of pixel-pairs. Results show that BRIEF easily outperforms other fast descriptors such as SURF in terms of speed, and it also outperforms them in terms of recognition rate in many benchmark datasets. Then, Rublee et al. (2011) propose ORB that is a very fast binary descriptor based on the BRIEF descriptor. The method produce a descriptor invariant to rotation and resistant to noise. ORB finds potential salient point locations on an image pyramid using the FAST edge detector (Rosten & Drummond, 2006) and picks the top set of points as salient points applying a Harris corner measure. The descriptor improves BRIEF comparing the intensity of patch-pair to form the binary string vector. Results demonstrate its improvements in performance and efficiency relative to other features. However, authors do not adequately address the scale invariance of the pyramid and they also do not explore per keypoint scale.

In the same year, Leutenegger et al. (2011) propose the BRISK descriptor. This method handles the problem of detecting, describing and matching image keypoints for cases without sufficient priori knowledge about the scene or viewpoint. Similar to ORB, this algorithm detect the interest regions using the FAST edge detector, but instead it uses a novel sampling pattern. It consists of sample points equally distributed on concentric circles centered around the salient point. Then, the algorithm determines orientation by computing local intensity gradients. Finally, the binary descriptor is a pairwise comparison of intensities. Experiments show this method offers a dramatically faster alternative at comparable matching performance than state-of-art interest region descriptors. Recently, Alahi et al. (2012) propose a novel descriptor based on human retina system, called FREAK. This method introduces a sampling retinal pattern that samples pairs of pixels

and then compares their intensities. The descriptor outperforms state-of-the-art keypoint descriptors while remaining simple faster with lower memory load which make it an excellent choice for mobile applications.

Most of these region descriptors have been recently compared in Z. Song and Klette (2013) and Wu and Lew (2013). Both evaluations agree that there is no a better interest region detector in all aspects, but that are dependent on the task. Wu and Lew (2013) reports that the FAST detector present highest repeatability score than other detectors, moreover and it had the least detection time cost per point. Their results also show that SIFT, BRISK, and FREAK are the best affine invariant descriptors, and the time complex showed the binary descriptors provide a very efficient description and matching. However, Z. Song and Klette (2013) report that SIFT is the best robustness descriptor with respect to rotation and scale changes, but its time issue has been confirmed again. In general, sparse key point approaches are characterized by their compact representation, i.e. there are fewer regions than image pixels. However, the key point detection algorithms do not guarantee repeatability for general object classification due to these interest regions are image dependent. This is a limitation in object categorization if the detector does not fire within the region occupied by the object. We defer the reader to these surveys for specific details about the evaluations.

2.1.2. Dense Sampling

This strategy proposes extracting features densely on the image or detection window and binding those features into a high-dimensional descriptor as shown in Fig.2.1b. The image is sampled using a grid to define local regions. Each region is represented using the intensity images, gradients or another appearance representation. The grid elements may be overlapping or not. Current algorithms use histograms of the computed feature to represent each local region. The local dense representation allows us to code visual features simultaneously with their location within the detection window. In addition, these representations avoid repeatability problems because they do not depend on search heuristics to find relevant regions. The most successful algorithms for dense representation are Haar wavelet coefficients (Papageorgiou & Poggio, 2000), Histogram of the oriented Gradients (HOG) (Dalal & Triggs, 2005), and Local Binary Patterns (LPB) (Ojala et al., 2000). In general, dense regular grid instead of interest points have shown work better for classification (Fei-Fei & Perona, 2005). In addition, this sampling method works better for discriminative classifiers

that can handle in an optimal way high dimensional feature vectors, such as Support Vector Machines (SVM) , Artificial Neural Networks (ANN) or boosting strategies (Tuzel et al., 2008). There are two variations related to the use of dense sampling that we describe below: multi-scale pyramid decomposition and spatial pyramidal decomposition.

Multi-scale pyramid decomposition comes from multi-resolution analysis using wavelet and space-scale analysis (Lindeberg, 1993). This sampling technique is popular in multi-scale analysis, where the objective is to extract features at various levels of detail by downsampling and smoothing the input image, as shown in Fig. 2.2a. In 2000, Papageorgiou and Poggio proposed a multi-scale detector with a set of wavelet features computed at fixed scales to represent samples of people, cars, and faces. The multi-scale detector uses the trained model and the same downsampling and smoothing procedure on a set of test images. In addition of the method, the authors submit the MIT dataset showing good performance of this method. Then, Dalal and Triggs (2005) introduce HOG feature that consists of a dense grid of gradient histograms with trilinear interpolation and local normalization. This feature shares properties with SIFT. The method uses a SVM with polynomial kernel to classify the multi-scaled windows sets. Results show an excellent performance level on the MIT pedestrian database and also on their new database, the INRIA Person. In 2009, Wang et al. propose the combination of the HOG feature and Local Binary Patterns (LBP). The LBP descriptor are invariant to strict monotonic changes in intensity, making this feature robust against changes in lighting. This combination aims to use the best from both features, while capturing shape and texture information. Maji et al. (2008) propose multi-scale histogram of oriented edge energy feature, similar to HOG, but with a simpler design and lower dimensionality. This method exploits the additivity property of the intersection kernel, in which the resulting decision function can be independently computed for each dimension. Results show on the one hand, an improvement in running time that is logarithmic in the number of support vectors, and on the other hand it produces classification rates significantly better than the linear SVM. In 2009, Dollar et al. propose using sums of a collection of low-level feature channels such as the CIELUV color space, gradients, and vote strengths for different HOG bins. The sums are computed using the integral image technique in order to accelerate the computation times while producing state-of-the-art results. Tuzel et al. (2008) introduce the covariance matrices descriptor as a way to combine different localized low-level features such as, intensity and gradients. They also use Riemannian manifolds, showing a substantial improvement over the INRIA Person dataset. In 2010, Felzenszwalb et al. use a similar

feature approach to Dalal and Triggs (2005) but representing the objects by a multi-resolution HOG over a deformable part-based model (DPM). Afterwards, Park et al. (2010) add scale as another latent variable to the DPM, using models specialized for the respective resolution ranges and a fixed HOG-like template at low resolutions. Results of using these sets of multi-resolution features show improvements in detection of large and small pedestrians. Recently, Dollar et al. (2014) improve the running time to compute a set of multi-resolution features that is the bottleneck of many modern detectors. They use an extrapolation method that is inexpensive as compared to the feature computation. Their results show that using this approach the DPM methods are completely suitable for real-time detector with fine sampled pyramids.

The pyramidal decomposition involves extracting features on smaller subregions of the image or detection window (Lazebnik et al., 2006; Bosch et al., 2007) as shown in Fig.2.2b. Unlike the dense representation, these subregions are generated by successively dividing each subregion until a maximum number of levels is reached. Thus, the use of the descriptor at various scales make it similar to the multi-scale pyramid decomposition. The spatial pyramidal decomposition aims to achieve both a global and local representation of the image. All features are collected in a high-dimensional vector. However, the decomposition levels should be limited in order to prevent over-fitting. In 2006, Lazebnik et al. introduce a simple and novel method for recognizing scene categories based on a spatial pyramid sampling. This technique consists in to partitionate the image into increasingly fine sub-regions and computing histograms of local features inside each sub-region. Aurtherors use a SIFT-like feature to describe these sub-regions and finally form a codebook based on the k-means clustering algorithm. Classification results show that this technique increases the performances for the datasets described in state-of-art methods. In the same way, Bosch et al. (2007) propose a method to classify images according to the object categories that they contain. They use a descriptor that represents local image shape and its spatial layout. These allows for the shape correspondence between two images can be measured by the distance between their descriptors by a spatial pyramid kernel. Their results significantly improves classification performance. In 2009, Yang et al. present a novel strategy based on sparse coding and a multi-scale spatial max pooling to generate discriminative codebooks from a spatial pyramids sampling. This approach reduces the complexity of SVMs in training and in testing. Another novelty factor is the use of sparse coding with appearance descriptors like SIFT features. Classification results show that this approach always significantly outperforms the linear with Spatial Pyramid Matching kernel on histograms. Recently, C. Zhang et

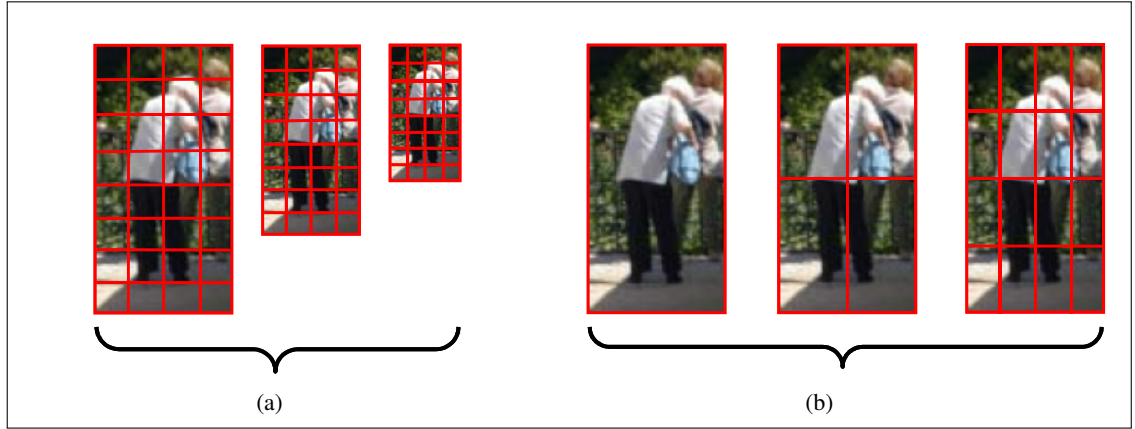


Figure 2.2. Examples of methods for sampling images. (a) Multi-scale pyramid decomposition sampling strategy: the image is downsampled and smoothed at every level and then sampled at smaller subregions. (b) Spatial pyramidal decomposition strategy: the image is sampled at smaller subregions by dividing the image until a maximum number of levels is achieved.

al. (2013) propose an image classification approach based on a spatial pyramid robust sparse coding technique. This sparse coding search a the maximum likelihood estimation by optimizing over the codebook and the feature coding parameters. This optimization allows the algorithm to filter out more outliers than traditional sparse coding methods. The visual codebooks formed using this sampling strategy generate more discriminative codebooks helps to improve the image classification performance. Results demonstrate the effectiveness of this method on the Scene 15 dataset and the Caltech 256 dataset.

2.2. Classification Methods

Building classification models is a crucial step in machine learning methods. These models are required in order to determine whether a set of features belongs to a specific class. There are two approaches to generating those models: discriminative and generative. Most of the detection applications based on sliding windows have been implemented using discriminative models due to their high level of performance, embedded ability to select relevant features, and ease of use. The most popular discriminative models are Support Vector Machines and Boosting.

Part of the appeal for Support Vector Machine is that non-linear decision boundaries can be learnt using the so called the kernel trick. Though this classification algorithm have faster training speed, the runtime complexity of a non linear SVM classifier is high. Boosted decision trees on the

other hand have faster classification speed but are significantly slower to train and the complexity of training can grow exponentially with the number of classes. Thus, linear kernel SVMs have become popular for real-time applications as they enjoy both faster training and classification speeds, with significantly less memory requirements than non-linear kernels due to the compact representation of the decision function (Maji et al., 2008).

2.2.1. Support Vectors Machines

Support Vector Machines algorithm (SVM) is a discriminative model that finds the separating hyperplane that maximizes the margin between two classes (Cortes & Vapnik, 1995). The output of the SVM does not provide posterior probabilities, however it can be calculated using the Platt method (Platt, 1999). SVM works either on the input features or a kernelized version of them. Let the training dataset comprise N input feature vectors x_1, \dots, x_N and their labels $y_1, \dots, y_N \in \{-1, 1\}$. The SVM maximize the objective function:

$$\min_{\mathbf{w}, \xi, b} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \right\}$$

subject to:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0,$$

where \mathbf{w} is the normal vector to the hyperplane, ξ_i is the slack variable which measures the degree of misclassification of the feature vector x_i , C is a cost constant to increase the penalization of errors, and x_i is the instance i in the train dataset. An important property of SVM is that the determination of the model parameters corresponds to a convex optimization problem. It ensures that any local solution is also a global optimum (Bishop et al., 2006). There are various kinds of kernels such as the Radial-basis kernel (RBF), Chi-square kernel, intersection kernels, etc. The selection of the kernel impacts the running time and the performance of the classifier. Previous studies have mainly used linear kernels because they perform better with large and sparse descriptors such as HOG and LBP (Dalal & Triggs, 2005; Felzenszwalb et al., 2009; Yang et al., 2009).

2.2.2. Boosting Classifiers

This is also a discriminative framework based on the idea of combining weak learners to get a strong classifier (Schapire, 1990). In 1995, Freund and Schapire proposed AdaBoost, which is an

adaptive version of boosting in which misclassified instances increase their weight during training. There are several versions of the boosting algorithm which were designed to deal with different problems and yield several kinds of outputs. In general, this framework is used to train cascades of weak classifiers. Despite the slow training time of the cascades, they represent an improvement in the run-time of the final detector.

Specifically, an AdaBoost algorithm builds a strong classifier as a linear combination of weak classifiers.

$$f(x) = \sum_{t=1}^T \alpha_t h_t(x)$$

where T is the number of training rounds or iterations; $h(t)$ is the weak classifier or can be thought of as features; and α_t is the weight of the feature t in the linear combination. The final or the strong classifier is defined as

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right)$$

2.3. People Detection

Single view approaches have received special attention from researchers as potentially useful for solving the people detection problem. Although there are multiple view approaches, most are still based on background subtraction frameworks with the inherent drawbacks of those approaches, as we discussed in Chapter 1. New methods include 3D cues, however, most of them work on single view camera configuration or require additional hardware. A specific analysis of each framework is provided in the next three subsections.

2.3.1. Single View Approaches

Significant progress has been reported in the use of machine learning in people detection. As we note in our literature review, there is a main group of single view approaches which use robust machine learning models and advanced features such as part-based models to cope with object variations or poses. In 2000, Papageorgiou and Poggio propose a general framework of object detection, including people, which use sliding windows, Haar wavelet features and a SVM model.

Then, Viola and Jones (2001) demonstrate that is possible to achieve real-time object detection by introducing integral images, a focus of attention mechanism. These same authors expand this work by including patterns of motion to improve people detection (Viola et al., 2005). In 2005, Dalal and Triggs introduce the histogram of oriented gradient features (HOG) to represent object categories. The method also uses a sliding-window approach to detect object instances, but in this case it was slower than the model used by Viola et al. (2005). Using the same method, Dalal et al. extend their work for detecting standing and moving people in videos based on orientated histograms of differential optical flow (HOF) (Dalal et al., 2006). They combine these motion-based descriptors with their HOG appearance descriptors, removing the false positive and increasing the overall detection performance in the state-of-art video datasets.

In 2007, Sabzmeydani and Mori propose a set of mid-level features to improve classification. The method uses a set of AdaBoost classifiers trained with regions of low-level features on the detection window to built a set of mid-level features that capture more information than the low-level feature sets. Then, Tuzel et al. (2008) present an algorithm to detect pedestrians in still images based on covariance descriptors (Porikli & Tuzel, 2006; Tuzel et al., 2006) and Riemannian manifolds. The descriptors form a d-dimensional nonsingular covariance matrices represented as a connected Riemannian manifold. Then, the classifier uses the geometry of the space to discriminate points lying on a the manifold. This approach can be combined with any boosting method. This algorithm achieves remarkable detection rates on the INRIA and DaimlerChrysler pedestrian data sets than state-of-the-art methods. In 2009, Wang et al. propose a people detection approach based on the combination of the HOG feature and Local Binary Patterns (LBP). The combination of both features allows the algorithm to capture the shape and the texture information simultaneously. This method also allow the detector to handle with occlusion using the SVM responses to infer the visible blocks in each window. The SVM score contributions of each block lets the algorithm to check if the contribution of the blocks is similar inside the detector window. When the SVM scores are inconsistent across the detector window, the algorithm employs a partial model trained using the portion of the window that is assumed to be visible. Schwartz et al. (2009) continues the idea of combining different types of features into high-dimensional feature vectors, over 170,000 dimensions. They propose pedestrian detector based on Partial Least Square Regression (PLSR) to reduce the dimensionality of the feature space. This technique allows any classifier to handle the problem of low amount of high-dimensional training data. The PLSR analysis is an efficient dimensionality

reduction technique that projects the data onto a lower dimensional subspace and preserves discriminative information. Results show that this method outperform state-of-the-art method on pedestrian datasets such as INRIA, DaimlerChrysler, and the ETHZ.

In the same year, Felzenszwalb et al. (2009) develop an algorithm based on combinations of multi-scale deformable part models using new methods for discriminative training with partially labeled data and cascade object detection. During training the algorithm uses a multiple instance learning scheme which applies a bootstrap strategy to select the best centered instances around the labeled bounding boxes in the train dataset. In 2014, Dollar et al. present a multiscale pedestrian detector for real time detection. This method attempts to solve the computational bottleneck of many modern detectors during the features extraction at every scale of a finely- sampled image pyramid. The algorithm approximates feature responses from nearby scales computed at a single scale. This allows the detector to accelerate the sampling of the image pyramid. The approximation yields a speedup over competing methods with only a minor loss in detection accuracy of about 1-2% on dataset described in the state-of-the-art. Recently, Pedersoli et al. (2014) propose a method focused on accelerate detection using computationally expensive hierarchical multiresolution part-based model (Felzenszwalb et al., 2009). The algorithm uses a coarse- to-fine strategy in order to speed up the search for pedestrians in an image. The method also includes reasoning about small objects generally missed by other detectors comparing scores of small objects where high-resolution features are not available with full-resolution detections. The algorithm runs over a graphics processing unit (GPU) to accelerate the feature computation. Results show that the method improves the running time of the part-based model in the context of driving assistance.

Despite the success of these detection approaches, they try in general to solve people detection problem using single view information and disregard spatial cues such as size and likely location. In some cases, the recall is degraded in order to increase precision. In 2011, Dollar et al. evaluate a set of pedestrian detectors trained in the same conditions and using the same datasets. They conclude that even though the algorithms have improved through the use of strong and sophisticated models, overall performance is still poor.

2.3.2. Multiple View Approaches

Multiple view object detection plays an important role in understanding and analyzing scenes captured by video cameras used today for wide area video surveillance applications. There are

challenging problems such as simple detection, tracking, camera location and topology discovery, and data fusion. Distinct issues arise according to the overlapping degree among the cameras. The overlapping multiple view camera systems are used in general for filtering disambiguations among the cameras (Aghajan & Cavallaro, 2009). Most of the earlier works related to people detection in overlapping multiple view approaches use background subtraction and tracking techniques to merge the subtracted images onto ground plane with overlapping fields of view. In 2006, Kim and Davis propose a multiple view people detector that uses the foreground blobs to delimit an occupied area on ground-plane homography. Then, a multiple view tracking algorithm discards false hypotheses along the sequence using people's appearance. The same strategy can be used to track basketball players using their jersey numbers (Delannay et al., 2009). The method uses this feature for tracking instead of appearance and estimate the bounding-box projection according to the subtracted blobs. Farhadi and Tabrizi (2008) present a method for activities recognition using multiple view. The algorithm builds models in terms of features that are stable at different viewpoints, but still discriminative. Transferring the activity models between views avoids for obtaining several examples of each activity in each view, impracticable in many cases. Although the overall results for activity recognition are still low, this works demonstrate the gain of using multiple views in this process. In 2010, Eshel and Moses, a ground-plane homography at various hypothetical head levels defines an occupancy map. A tracking algorithm uses feature location and position of the head projected to the floor. A total number of eight cameras with top-view alignment yield the best performance. A year later, Han et al. (2011) propose a tracking method for fusing information from multiple sensors. The algorithm estimates the current tracker state by using a mixture of sequential Bayesian filters and changing dynamically different level of contribution of each camera to estimate a more reliable posterior. Results show that this tracking method outperform other sensor fusion techniques based on probabilistic tracking such as the Kalman Filter and its variants, and particle filter.

In 2012, Yildiz and Akgul propose a method for multi-camera multi-person tracking. The method uses constraints from the epipolar and projective geometries. As in Han et al. (2011), this algorithm computes the projection of a probability mask of the object positions on the ground. It also includes a voting method based on the employment of the integral images to make this computation very fast. A tracking algorithm based on methods like the Kalman Filter locate people along the time. One of the drawbacks is that the algorithm turns inaccurate when time intervals are too long. This drawback is also present in the previous methods on multi-camera people tracking. Then,

Evans et al. (2013) present an people detection algorithm in multiple views based on projecting foreground mask into a common coordinate system. This method attempts to handle suppression of false detections and automatically estimate the size of the objects under tracking. The preliminary results show uses of this method in situations of mixed pedestrians and vehicles. Recently, Liem and Gavrilu (2014) introduce a method to multi-person tracking in overlapping cameras configurations. The method uses two-step that jointly estimates the person position and the track assignment and keeps this assignation problem tractable. During detection the algorithm evaluates the similarity between a person at a particular position. This information lets to active a track based on cues such as appearance and motion. Results demonstrate that the algorithm outperforms the state-of-the-art on four challenging multi-person datasets.

In general these methods rely on a segmentation procedure and a tracking algorithm for detection. In addition, these methods must infer the size instead of searching for it. Heads detection is a special case of people detection. Important conclusions from works of people detections claims that detecting the head helps avoid occlusions in crowded environments due to head is the least occluded part of the body during this conditions, and it is also less deformable than the rest of the body. (Dalal & Triggs, 2005; Eshel & Moses, 2010; Ali & Dailey, 2012). There are examples of heads detectors incorporated as part-based detector into more complex detections systems in order helps to find the complete body or the other body parts, and improve the detection (Zeng & Ma, 2010; Ali & Dailey, 2012; Xie et al., 2012; Nghiem et al., 2012; Chang et al., 2013; Hayashi et al., 2013). Our method uses this principle and is focused on head detection without the use of tracking.

2.3.3. Related Approaches for Detection

In general, the 3D recognition using 2D images is a difficult task due to the infinite number of viewpoints and varied lighting conditions (Poggio & Edelman, 1990; Ikeuchi & Kanade, 1988). Initial attempts to represent 3D objects were based on the psycho-physic premise that a 3D structure can explain all of the changes in appearance that arise from viewpoint changes or aspects of the object (Mundy, 2006). In 1979, Koenderink and Doorn use the aspect-graph to establish the relationship to the topological appearance of the object. A node of the graph represents adjacent object views, and an edge rises from the transition in the graph that relates to the vertices, edges and faces of the projected object (Cyr & Kimia, 2004), as is shown in Fig. 2.3a. Following the idea of appearances, Ikeuchi and Kanade (1988) proposed an object recognition system based on grouping

views or aspects using similar features. The algorithm extracts features such as object faces and edges under various viewer directions sampled on Gaussian sphere. These features describe each of the object aspects. Then, the algorithm uses the features to group similar aspects and build an interpretation tree which is in turn used for recognition. First, this tree classifies the unknown view into the correct aspect. Second, it determines the actual position of the unknown view in that aspect. The final performance depends on the chosen features set.

In 1998, Pontil and Verri used aspects to represent and classify objects with an SVM classifier. Each object class is represented by a set of aspects from various viewpoints. A set of 36 images per class is used to train the SVM classifier, which learns the margins using one-vs-one strategy to separate the selected classes. During the test, the SVM model evaluates the class for a new and unseen instance from a random viewpoint. The recognition is independent of the object pose. In 2004, Cyr and Kimia proposed a method for recognizing 3D objects from 2D images using an aspect-graph structure and a shape similarity measure. A viewing sphere is sampled at regular interval of five degrees as shown in Fig.2.3b. Next, an iterative procedure combines views into aspects with a prototype representing each aspect. During recognition, a new instance is compared against the set of prototypes using the same dissimilarity measure. Following this idea, T.-M. Su et al. (2006) present a method for recognizing 3D objects from 2D images based on Fourier descriptors of sampled points over the object contour. The algorithm computes point-to-point lengths as features and similarity metrics to establish the canonic views as the aspects. The experiments using human postures show the effectiveness of the method to represent rigid and non-rigid objects. However, the computation time required to run the method is high. Toshev et al. (n.d.) propose a method for recognizing moving objects in videos based on silhouette features in order to make the method independent of the objects appearance. The algorithm extracts the object silhouettes from video by segmentation of successive frames. Finally, it matches these silhouettes to the 3D model silhouettes in a synthetic dataset. The method achieves promising experimental results using purely shape-based matching scheme driven by synthetic 3D models.

Following the idea of aspect-graphs, Ulrich et al. (2012) present a 3D object recognition approach based on hierarchical model generated using only the geometrical information of a 3D CAD models of the objects. The algorithm generates the hierarchical model projecting a 3D CAD model on images and taking into account the scale-space effects. Finally, it refines the 3D pose using

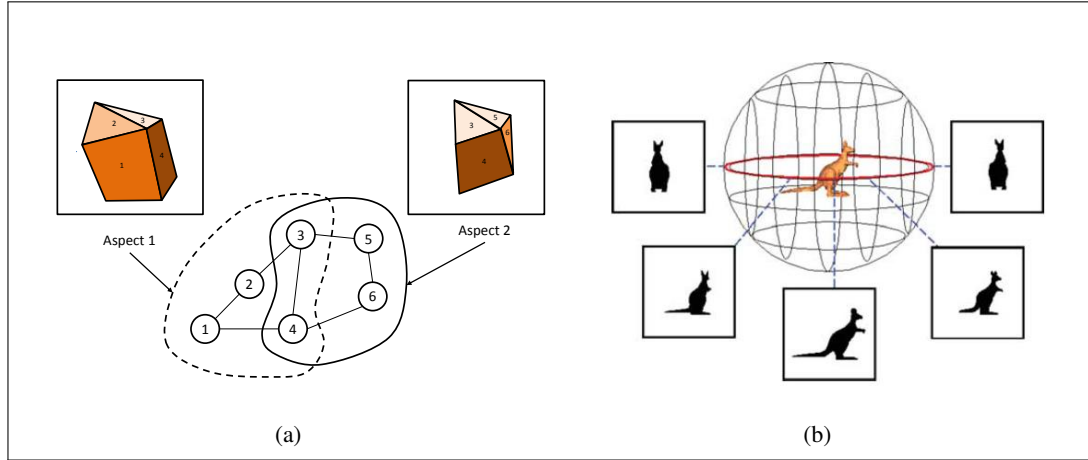


Figure 2.3. Examples of an aspect-graph. (a) Two views or aspects of a polyhedral object. These aspects correspond to projected faces, which form a graph according with their adjacency, *i.e.*, faces 1, 2, and 4 in the Aspect 1 appear connected because they are adjacent. (b) The viewing sphere sampled at regular intervals. During sampling, images are captured every five degrees. An iterative procedure combines views into aspects using prototypes which represent each aspect.

a least-squares adjustment that minimizes geometric distances in the image. Due to computation efficiency, authors must limit the pose range depending on the application. The major limitation is that the pose range should not contain any degenerated views of the object, like a side view. Recently, Atmosukarto and Shapiro (2013) propose a object retrieval method based on selecting the salient 2D views to describe 3D objects. This method uses a silhouette-based similarity measures to discriminate among different synthetic 3D objects. The algorithm computes this measure over multiple salient points on the silhouette of the object. Retrieval experiments show that the use of salient views are promising.

A similar way to 3D object recognition consists in to perform the matching of invariant features. However, this simple strategy may fail when objects present significant intra-class variation. In Rothganger et al. (2006), a novel representation for 3D objects is presented based on local affine-invariant image descriptors and multiple view spatial constraints. The algorithm exploits the idea that smooth surfaces are always locally planar. Thus, the matching and recognition are possible using photometric and geometric consistency constraints. A disadvantage of this approach is its poor performance on texture images. In Ferrari et al. (2006), a similar method is presented based on the relationships between multiple model views. It enforces global geometric constraints in order to achieve 3D reconstruction from multiple views for recognizing single objects. A disadvantage is

its poor performance on non-textured images and uniform objects. There are also other approaches to 3D object category classification which use 3D data such as points, meshes and CAD models. In these methods, the reconstructed 3D object serves as a query and is matched against the shape of a collection of 3D objects. Some approaches are: multiple view range data and CAD model (Li et al., 2006); range data with local feature histograms (Hetzl et al., 2001); 3D shape descriptors based on a spherical harmonic representation (Kazhdan et al., 2003); and recognition using only a part of the object based on a *thesaurus* (Ferreira et al., 2010). For a complete list, see Bustos et al. (2005). However, these methods are not used in practice to recognize real world objects in cluttered scenes because they are not able to recognize the underlying structure of the object (Savarese & Fei-Fei, 2010). Recently, Zia et al. (2013) introduce a recognition method based on 3D geometric object class representations. The algorithm uses selected vertices from a 3D CAD models dataset to train geometrical representation. The a local shape representation allows the algorithm to match the geometric model with real-world images at different viewpoints. Each vertex in the 3D model have associated a detector that identifies this part in the images. Finally, during recognition the model finds an instance in the 3D model that best explains the observed image. Results in 3D pose standard benchmark datasets show the ability of the method to accurately localize objects and their geometric parts in 2D.

Few years ago, Helmer and Lowe (2010) proposed a combined approach of sliding-windows detector and depth images. The results demonstrate that using appearance or shape information is not enough for detection, but that using spatial information improves the performance of a multi-scale detector. They start by locating the object using a multi-scale sliding-windows classifier. These detections are filtered applying a depth image from stereo data to improve object location, reducing false positives and increasing the scores for true positives. However, this approach focuses on mobile robots and requires additional hardware to create spatial information. In 2011, Benenson et al. proposed a quick method for computing stixel words, which allow researchers to model the world locally as flat sticks rising vertically above the ground. The stixels increase detection performance by reducing the number of candidate windows required to detect people. This approach fits well with robotic platforms because it requires stereo cameras to compute a ground plane estimate based on disparity maps.

The psycho-physic idea of recognizing objects using aspect-graph-based models presents powerful advantages. However, these models present several practical disadvantages: *i*) the size of the aspect-graph grows rapidly with the topological transitions required for object recognition, which implies that the aspect-graph becomes application specific (Mundy, 2006); *ii*) the scale required to determine the relevant transitions in accordance with the object topology (Mundy, 2006); and, *iii*) the complexity of generating the aspects and the storage and search requirements, which are impractical for objects of modest complexity (Cyr & Kimia, 2001). However, the geometric 3D reasoning has received renewed attention recently. In this context, the level of geometric detail is typically limited to qualitative or coarse-grained quantitative representations, which can recover geometrically accurately object hypotheses than naive 2D bounding boxes. Nonetheless, the object class detectors are tuned towards robust 2D matching rather than accurate 3D pose estimation, encouraged by 2D bounding box-based benchmarks such as Pascal VOC (Zia et al., 2013). Progress has been made in this field using 3D object classification and detection, especially in regard to linking features among views in a discriminative learning framework to create multiple view models of objects. However, they are still single view detectors and do not solve the problem of combined multiple view detections or including 3D information on the location and the size of the objects.

2.4. Non-Maximal Suppression

Non-maximal suppression (NMS) is a critical procedure in computer vision algorithms that selects only one representative point from a set of interest points. The output of an object detection algorithm is a set of multiple detections around the most confident one. These detections are filtered using a non-maximal suppression (NMS) algorithm. The most popular alternatives are Mean Shift (Fukunaga & Hostetler, 1975; Y. Cheng, 1995; Comaniciu & Meer, 2002) and Local Maximum Searching (Felzenszwalb et al., 2009; Dollar et al., 2009).

The Mean Shift algorithm or the Parzen window technique is a non-parametric procedure used to locate the maxima of a density function from discrete data samples (Fukunaga & Hostetler, 1975). Dalal (2006) propose a weighted Mean Shift procedure to suppress the multiple detections provided by their people detection algorithm. This procedure requires a set of detections $\mathbf{y}_i, i = 1 \dots n$ defined by a location in the image and a scale. Each detection is taken as a point with an associated symmetric positive definite 3×3 bandwidth covariance matrix \mathbf{H}_i to define the smoothing width

for the detected position. The algorithm fuses the overlapped detections to represent the n points as local modes. Dalal also assume the smoothing kernel as Gaussian such that the weighted kernel density estimate at a point is given by

$$\hat{f}(\mathbf{y}) = \frac{1}{n(2\pi)^{3/2}} \sum_{i=1}^n |\mathbf{H}_i|^{-1/2} t(\varpi_i) \exp\left(-\frac{D^2[\mathbf{y}, \mathbf{y}_i, \mathbf{H}_i]}{2}\right)$$

where

$$D^2[\mathbf{y}, \mathbf{y}_i, \mathbf{H}_i] \equiv (\mathbf{y} - \mathbf{y}_i)^T \mathbf{H}_i^{-1} (\mathbf{y} - \mathbf{y}_i)$$

is the Mahalanobis distance between \mathbf{y} and \mathbf{y}_i , and the term $t(\varpi_i)$ provides the weights for each detection assigned by the classifier. The local mode was iteratively estimated computing

$$\mathbf{y}_m = \mathbf{H}_h(\mathbf{y}_m) \left[\sum_{i=1}^n \varpi_i(\mathbf{y}_i) \mathbf{H}_i^{-1} \mathbf{y}_i \right]$$

from \mathbf{y}_i until \mathbf{y}_m stops changing. The term \mathbf{H}_g is the weighted harmonic mean of the covariance matrices \mathbf{H}_i . For more details about the weighted Mean Shift algorithm, see Dalal (2006).

In local maximum searching, the detections with the highest confidence values prevail. In (Dollar et al., 2009), the authors propose a simplified NMS procedure which suppresses the less confident of every pair of detection sufficiently overlapped according to the PASCAL criterion (Everingham et al., 2006). In this way, the only parameter required is the overlap threshold. Felzenszwalb et al. (2009); Dollar et al. (2009) prefer the local searching for NMS to the Mean Shift method proposed in Dalal and Triggs (2005) because its efficient running time during detection, and the fewer parameters to set up than the procedure proposed by (Dalal, 2006), such as the transformation function, the threshold of the transformation function, and the bandwidth.

2.5. Focus-of-Attention

Detection is a complex task in computer vision that requires efficient use of computer and sensory resources. Modern detection algorithms based on sliding-window include techniques for selecting a subset of regions in the images in order to improve the performance and speed of the detection. The process of directing the attention to these interest regions is called focus-of-attention

(Itti, 2000). Though attention mechanisms generate overhead, they pay off due to the complexity of detection, *i.e.*, the more general the detector, the more important the pre-selection of the region of interest (Papageorgiou & Poggio, 2000; Rutishauser et al., 2004; Frintrop et al., 2005).

In 1999, Itti and Koch present an algorithm inspired by biological visual systems to compute the focus-of-attention in images as a predictor of human fixation in images. This framework combines multi-scaled center surround feature maps and computes a single saliency map containing the main regions of the images likely to contain interesting objects. For general cases, the algorithm uses a simple and non-specific normalization for combining the feature channels and generates the saliency map. Nonetheless, this procedure can also include a learning stage by weighting the feature channels in order to detect specific object targets classes. Finally, in both scenarios a winner-takes-all (WTA) strategy is used to select the most salient objects. In Papageorgiou and Poggio (2000), the algorithm includes this visual attention strategy as preprocessing stage to enhance the processing speed of the system. The mechanism allows the algorithm to focus only on areas of the image that are likely to contain people. In 2005, Frintrop et al. present VOCUS, a top-down framework that is used to perform a goal-directed search of salient objects in images. The algorithm first learns a set of weights to combine the bottom-up features from a full labeled training dataset as in (Itti & Koch, 1999), and then computes a global saliency map from the difference between the excitation and inhibition maps. Finally, the algorithm uses A WTA strategy to detect the salient objects in the image. The top-down strategy enhanced the selection of salient regions and improved recognition performance. In (Davis et al., 2007) propose a focus-of-attention procedure for surveillance. The method builds an adaptive model of the scene based on the local human activity at discrete locations. A motion history images technique and a classification algorithm allows for the temporal signature of translating objects represented as blobs be captured and separated from noise. Finally, an activity map summarizes these temporal signatures. The algorithm lets to maximize the opportunity of observing human activity. In the same year, Hou and Zhang (2007) propose a simple method for visual saliency detection based on the analysis of the log-spectrum of an input image. The algorithm extracts the spectral residual of an image in the spectral domain in order to construct its corresponding saliency map in the spatial domain. This method works independently of features, object classes or prior knowledge. However, it still depends on the image scale and threshold settings to produce the saliency map. The spectral residual handle the issue of weighting features from different channels. Results show the low computational load, and the effectiveness of this approach in real images and

psychological patterns. Liu et al. (2011) introduce a supervised method for salient object detection based on a set of novel features and Conditional Random Fields (CRF). The features include multi-scale contrast, center-surround histograms, the color spatial distribution, and dynamic salient features used for detect salient object in image sequences. These features allows for the salient object be described locally and globally. The CRF learns to combine the feature channels. Results show that CRF with all three features produces the best result.

Earlier attention mechanisms generally rely on the information contained in the images and require a supervised procedure to learn salient object. In 2012, Shen and Wu present a novel salient object detector based on a unified model. Beside the typical use of the bottom-up and the top-down relations, this model incorporates links between low-level features and higher-level knowledge to guide the detection. The model represents images as a low-rank matrix plus sparse noises in a certain feature space. On the one hand, the low-rank matrix explains the non-salient regions. On the other hand, the sparse noises explains the salient regions. The model also includes higher-level knowledge such as location, semantic and color, that acts as priors to generate the saliency map. Results show that the method achieves comparable performance to the existing methods even without the help from high-level knowledge, and outperform the state-of-the-art when it uses the high-level priors. Recently, Siva et al. (2013) propose an unsupervised method for salient objects detection based on a probabilistic formulation. The algorithm learns the most interesting patches likely to correspond to an object from a large corpus of unlabeled images. A sampling procedure proposes the object location in the saliency map, because in this unsupervised scenario the algorithm does not previously know what a person will find salient. The sampling occurs from similar images according to the GIST descriptor and color similarities. Results show that using only a single object location proposal per image the algorithm is able to select an object in the PASCAL VOC 2007 dataset. Authors show that the method can also be used as a simple unsupervised approach to the weakly supervised annotation problem. We defer the reader to Borji et al. (2012) and Frintrop et al. (2010), for a detailed survey and benchmark of salient object detectors.

In Viola and Jones (2001), the authors proposed a variant of the attention mechanisms using a cascade of classifiers to increase detection performance while radically reducing computation time during detection. The first classifiers in the attentional cascade are weak and fast, allowing the algorithm to reject negative windows in the earlier stages of the cascade. The last classifiers in the

cascade are very accurate, but computationally expensive. Though the attentional cascade is successful, its main drawback is the training complexity when this technique is used in large datasets (Pedersoli et al., 2010). In 2010, Felzenszwalb et al. improved their part-based model detector using an attentional cascade which uses thresholds to prune partial hypotheses computed from a simpler version of them. This idea is refined in Pedersoli et al. (2010), where authors used a coarse-to-fine strategy with a part-based model to select the most admissible regions of the image. They proposed an algorithm which uses a set of features at different resolutions in the same classifier, and used location refinement to speed up the image scan. This approach allows the algorithm to avoid using thresholds in the cascade, which may degrade the quality and performance of the detector. In 2012, Alexe et al. introduce an objectness measure that quantifies how likely it is for an image window to contain an object of any class. This is a generic measure. It combines several image cues by a Bayesian framework, including an innovative cue to measure the closed boundary characteristic. Results show that the new image cue outperforms traditional saliency features and the combined objectness measure performs better than any cue by itself. Results also demonstrate that the method greatly reduces the number of windows evaluated and can act as a focus of attention mechanism. Ali Shah et al. (2013) present an algorithm for object detection based on the objectness measure (Alexe et al., 2012). This method does the automatic object detection by correctly estimate the number of required windows, independently of the object classes present in the scene. Experiments on PASCAL VOC 07 dataset reveal that the algorithm outperforms prior works and provides a more accurate estimation of the required number of windows for an input image. Recently, M.-M. Cheng et al. (2014) present BING, a Binarized Normed Gradients descriptor for objectness estimation. The method uses the norm of gradients on a 8x8 window into a 64D feature to describe it. The algorithm uses this descriptors to train a generic objectness measure. Then, the objectness estimation requires only a few atomic operations such as *add*, *bitwise*, *shifts* to calculate sets of binary patterns. Results shows that this approach efficiently runs at 300fps on a single laptop CPU and generates a small set of category-independent, high quality object windows. The method achieves high detection rate on PASCAL VOC 2007 dataset. Although the promising results, the method presents a few limitations related to the small size and location of the predicted bounding boxes for some objects categories.

2.6. Discussion

As we noted in the literature review, people detection have achieved significant progress due in part to the use of powerful machine learning models, new more informative visual features, and part-based models which cope with the objects variability. Most of the detection approaches run a sliding-window which creates a densely sampled image pyramid in order to perform the multi-scale detection (Papageorgiou & Poggio, 2000; Dalal & Triggs, 2005; Felzenszwalb et al., 2010; Dollar et al., 2014). A common denominator of these detectors is that they mainly rely on statistical learning methods that exploit image-intensity information to capture object appearance features. Their goal is to uncover visual spaces where visual similarities carry enough information to achieve robust visual recognition. As a relevant limitation, appearance-based approaches do not incorporate relevant geometric information that can provide useful and relevant spatial cues such as the real size of the object to be detected, depth, and spatial likely appearance location. These spatial cues help reduce the number of false hypotheses (Helmer & Lowe, 2010; Benenson et al., 2011). Some notable exceptions with promising results are Salas and Tomasi (2011); Spinello and Arras (2011); Espinace et al. (2013); however, these approaches require additional hardware to recover the spatial information. Although the latest and most successful approaches use part-based models (Felzenszwalb et al., 2010; Pedersoli et al., 2010), there are situations in which occlusion prevents these models from working properly. Heads detection is a special case of people detection. Important conclusions from works of people detections claims that detecting the head helps avoid occlusions in crowded environments. The head is the least occluded part of the body during this conditions, and it is also less deformable than the rest of the body. (Dalal & Triggs, 2005; Eshel & Moses, 2010; Ali & Dailey, 2012). There are examples of heads detectors incorporated as part-based detector into more complex detections systems in order helps to find the complete body or the other body parts, and improve the detection (Zeng & Ma, 2010; Ali & Dailey, 2012; Xie et al., 2012; Nghiem et al., 2012; Chang et al., 2013; Hayashi et al., 2013). We based our method on this assumption because this is an important design issue when detecting people in indoor environments such as classrooms or elevators where a body could be occluded by other persons or objects in the environment.

Models based on the psycho-physic idea of modeling different object aspects from multiple viewpoints present relevant advantages for detecting objects which are related to similarities for representing objects in the human brain (Stone, 1999; Tarr & Kriegman, 2001). Most of these

algorithms are trained using data from various viewpoints, however, detection occurs still on single camera configurations (Cyr & Kimia, 2004; Thomas et al., 2006; Savarese & Fei-Fei, 2007; Kushal et al., 2007; H. Su et al., 2009). There are also practical limitations such as the scale required to identify relevant transitions and the complexity of the graphs, which suggests that other models should be used to recognize the underlying structure of the object in real world problems (Savarese & Fei-Fei, 2010).

The results of new people detection techniques are promising and highlight new paths for research. However, the idea of how to combine information from various observation points has not yet enough maturity. Detection using one camera is suitable when there is mild occlusion, but in situation of heavy occlusion multiple views contributes to improve the final detection (Mittal & Davis, 2003; Khan & Shah, 2006; Eshel & Moses, 2010; Liem & Gavrilu, 2013). There are geometric techniques based on epipolar geometry and camera calibration that allow us to establish relationships across views in a camera system, providing useful ways to combine and integrate this information (Hartley & Zisserman, 2003; Szeliski, 2010). However, there are practical constraints when appearance of the target object change drastically among viewpoints (Lowe, 2004; Mikolajczyk & Schmid, 2005). These facts suggest that even if there are multiple view approaches based on matching and multiple view geometry to combine detection in multiple observing points, an alternative framework may be required to generate combined and corresponding projections of the object without using standard matching techniques. We also note that aspect-based representation fits well with multiple view environments where aspects of people are acquired from various viewpoints simultaneously. Further, using this representation in a multiple view framework might include enriched appearance information and 3D cues, which help to locate people in the scene or to filter out regions where people is not likely to appear and to generate detections according with the size of people.

Wide-baseline multiple view frameworks allow for the generation of spatial information similar than ranger sensors or stereo imaging does. These multiple view frameworks use camera parameters and geometric constraints between viewpoints to execute spatial reconstruction of the environment and the objects within it. Though the analysis of the 3D space using multiple view data generates significant overhead for detection algorithms, a spatial focus might drastically reduce the number of hypotheses required to analyze and improve detection performances as people detection in 2D

images (Itti & Koch, 1999; Viola & Jones, 2004; Frintrop et al., 2010; Alexe et al., 2012). In the next Chapter, we consider these issues and introduce our approach.

Chapter 3. PROPOSED APPROACH

Detection methods aim to uncover locations in images where visual similarities carry enough information to achieve robust visual recognition. Current methods for detecting people in images run a sliding-window over an image horizontally and vertically. Each window is classified as people/no-people and a suppression procedure prevents to multiple detections. The significant progress in the detection is due in part to the use of: powerful machine learning models (Cortes & Vapnik, 1995; Freund & Schapire, 1995; Breiman, 2001; Bengio, 2009; Lee et al., 2009); new more informative visual features (Lowe, 2004; Dalal & Triggs, 2005; Ojala et al., 2000; Maji et al., 2008; Wang et al., 2009; Calonder et al., 2010; Rublee et al., 2011; Dollar et al., 2014), ;and part-based models which cope with the object variability (Felzenszwalb et al., 2010; Pedersoli et al., 2010; Girshick et al., 2014; Dean et al., 2013). Though there have been some improvements, the overall performance is still poor (Dollar et al., 2011). As a limitation, appearance-based approaches do not incorporate relevant geometric information that can provide useful and relevant spatial cues such as the real size of the object to be detected, depth, and spatial likely appearance location.

In a multiple view scenario, detections supplied by a detection algorithm can be combined in order to discard false detections and enhance the overall detection performance. This combination of information allows false detections in one camera to be discarded or missed detections to be added in the final set of detections (Mittal & Davis, 2003; Khan & Shah, 2006; Eshel & Moses, 2010; Ali & Dailey, 2012; Liem & Gavrila, 2013). Multiple view geometry provides to us of rules for combining detections among cameras in a wide-baseline stereo configurations (Hartley & Zisserman, 2003; Szeliski, 2010). Nonetheless, there are practical constraints when appearance of the target object class change drastically among viewpoints that make the matching task is not trivial. For example in head detection, it means that features between views are dissimilar when they take opposite views of the head, or when one of the cameras fails to detect them (Lowe, 2004; Mikolajczyk & Schmid, 2005).

The geometric 3D reasoning has received renewed attention recently, using 3D object classification and detection, especially in regard to linking features among views in a discriminative learning framework to create multiple view models of objects (Zia et al., 2013). In this sense, combining information in a multiple view camera configuration is close to the psycho-physic approach for representing the 3D structure by aspects of the object target class (Pontil & Verri, 1998; Cyr

& Kimia, 2004; Mundy, 2006). An aspect-graph establishes the relationship to the topological appearance of the object. In this graph, a node represents adjacent object views, and an edge arises from the transition in the graph that relates to the vertices. The psycho-physic idea of recognizing objects using aspect-graph-based models presents powerful advantages, but also several practical limitations such as: the size of the aspect-graph (Mundy, 2006); the scale required to determine the relevant transitions (Mundy, 2006); and, the complexity of generating the aspects and the storage and search requirements (Cyr & Kimia, 2001).

These facts suggest that even if there are multiple view approaches based on matching and multiple view geometry to combine detection in multiple observing points, an alternative framework may be required to generate combined and corresponding projections of the object without using standard matching techniques. We propose a 3D extension of the single view sliding-window approach to multiple views configurations. In our approach, we run a volume element through three directions (X, Y, Z) of the world frame instead of running a sliding-window in the 2D domain. We call to this volume a *sliding-box*. This sliding-box defines sets of projections on images according to its size and location. We use the set of projections of this *sliding-box* on 2D images to decide if this space location belongs to an object target class, *e.g.*, people/no-people, head/no-head.

The main idea of our method is to analyze only the portion of the images in which the sliding-box is projected. This means that the *sliding-box* is geometrically projected according to its size and shape onto the images in the camera system in order to place projected bounding-boxes on each image, as shown in Fig.1.2. The set of projections forms a collection of aspects (Cyr & Kimia, 2004), as shown in Fig.3.1a. We refer to those prototypical views or templates of an object that are similar to each other as an aspect. In this way, we simultaneously consider all of the information in the multiple view camera system to enhance detection as compared to an approach that uses only a single view. We replace the matching task by a classifier that learns the correct order and alignment of these aspects. A set of features vectors summarize measures of each projection, and serves as inputs vectors to a multiple view classifier. We make a decision regarding the joint data build up using all of the projected windows. The proposed approach has five main steps: spatial focus-of-attention, multiple view projection, feature extraction, classification and non-maximal suppression, as shown Fig. 3.1b. Detailed descriptions of these steps will be provided in the five sections that follow.

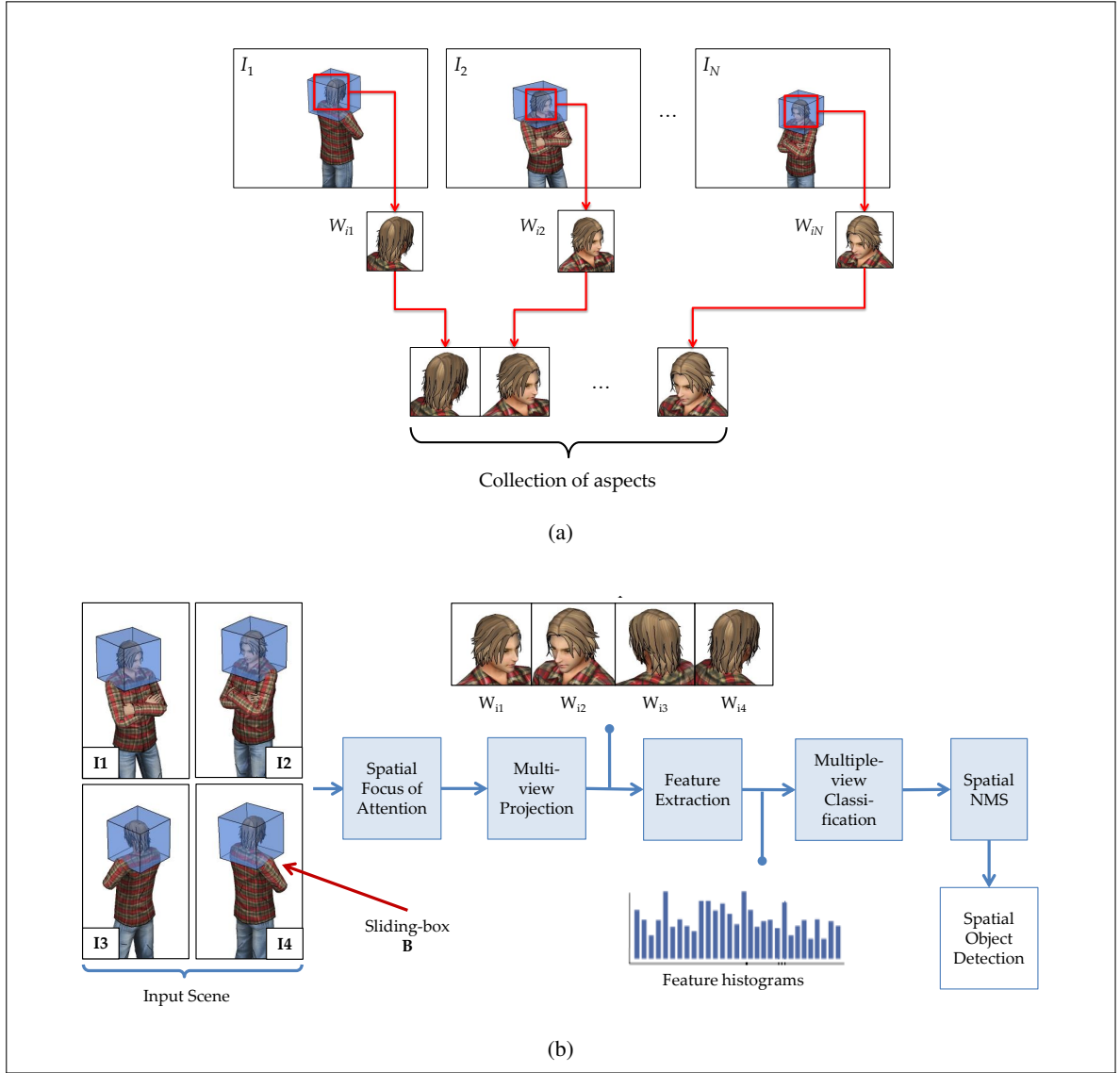


Figure 3.1. (a) Example of the aspects collected from multiple view projections using our proposed approach 1. The set of projections form a collection of aspects of the head. We use this projection strategy to collect all of the information available from all of the view-points in the camera system. (b) Block diagram of the proposed method. Our approach includes five main steps: spatial focus-of-attention, multi-view projection, feature extraction, classification and non-maximal suppression. In the example, we use $N = 4$ cameras. The algorithm begins with an input scene composed of I_1, \dots, I_4 images. Then, the spatial focus-of-attention reduce the number of head hypotheses. Next, the algorithm computes the projections of the box B_i onto the image I_j as W_{ij} , forming a collection of aspects. Afterwards, we extract a set of features for each projection W_{ij} and apply the model for sequences. Finally, a spatial NMS procedure allows us to eliminate multiple detections.

3.1. Spatial Focus-of-attention

A relevant limitation to apply directly the proposed sliding-box method is the demanding computational complexity to project and process the huge amount of boxes generated in the 3D space. Modern detection algorithms include attention mechanisms for selecting the most salient regions in the images (Itti, 2000). These mechanisms allow the detection algorithm to improve the performance and speed during detection. The attention mechanisms generate overhead, however, they pay off due to the complexity of detection (Papageorgiou & Poggio, 2000; Rutishauser et al., 2004; Frintrop et al., 2005). Earlier focus-of-attention algorithms predict what humans will label as interesting in an image by combining low-level feature channels using a bottom-up strategy (Itti, 2000; Hou, 2007). Then, the use of a top-down framework and novel saliency features allows these attention algorithms to detect salient object in images (Frintrop et al., 2005; Montabone & Soto, 2010; Liu et al., 2011; Shen & Wu, 2012; Siva et al., 2013). Recently, object saliency approaches use supervised learning to generate an objectness measure that allows the detection algorithm to select the most likely windows to contain the object target class. These recent attention mechanisms provide an output similar that the standard saliency maps, however, they also work as specialized object detectors. (Marchesotti et al., 2009; Rahtu et al., 2011; Alexe et al., 2012; Ali Shah et al., 2013).

In order to reduce the burden during detection, we apply a pre-processing procedure based on an attention mechanism in the spatial domain. In our method, we define the sliding-box movements in a spatial mesh grid equally spaced in which all of the sliding-box positions are previously known. The spatial mesh generate a large number of boxes, most of them located in places without presence of the object target class. There is also a trade-off between the step size and detection accuracy. Rough grids with large steps reduce the total number of boxes to analyze but are less accurate. On the other hand, fine grids with small steps generate an extremely large field for analysis but increase accuracy. Fortunately, we further reduce the search space by applying this pre-processing step that consist of using a salient object detector and multiple view geometry to generate a focus-of-attention in 3D space. Thus, the sliding-box runs only across these set salient spatial points, as we show in Fig. 3.2. There are coarse-to-fine strategies for finding focus-of-attention in images which progressively reduce the burden of the algorithm during detection and enhance the final detections (Pedersoli et al., 2010). Nonetheless, the sliding-window must still be run at the first levels of the

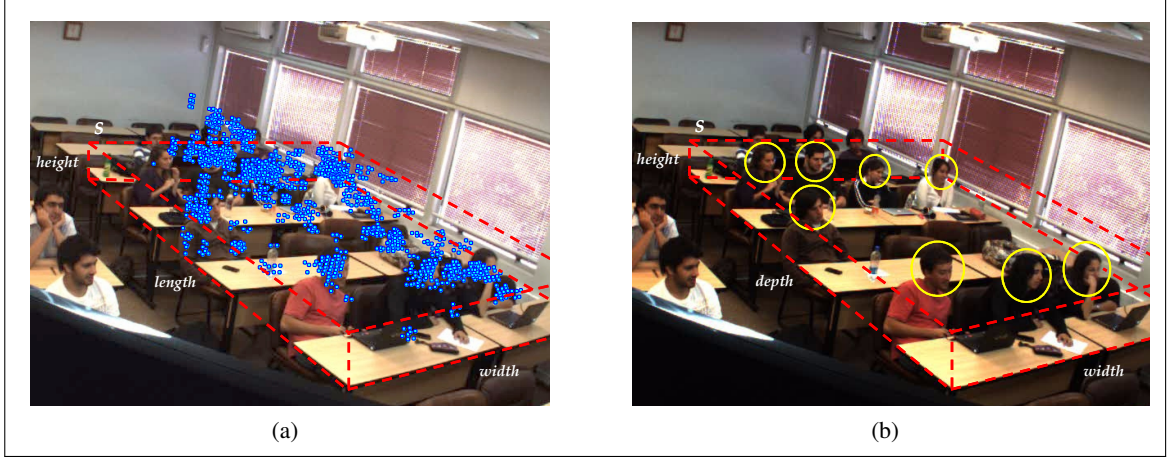


Figure 3.2. Explanation of the focus-of-attention procedure. (a) Blue dots show the potential head positions detected by our focus-of-attention procedure. This process provides hypotheses with more likely location of heads within the region of interest S and helps us to drastically filter the spatial detections. (b) Yellow circles shows ground-truth heads examples within the subspace S .

feature pyramid. We cannot directly apply these kinds of approaches because they do not reduce the burden of running the sliding-box through the 3D space.

In order to compute the focus-of-attention, first we run a salient object detector based on a standard sliding-window detector as specialized region detector in the 2D domain. This procedure allows us to find head hypotheses in the images. Let \mathbf{h}_{ij} be the image coordinates for the centroid of the head hypothesis i in the image j . Next, we match the hypothesis \mathbf{h}_{ij} with the set of hypotheses \mathbf{h}_{rs} closer to the epipolar line $\mathbf{l}_s = \mathbf{F}_{js} \cdot \mathbf{h}_{ij}$ in the image s , where \mathbf{F}_{js} is the fundamental matrix computed with camera matrices. Then, the spatial location $\hat{\mathbf{M}}_i$ for each pair of matched hypotheses \mathbf{h}_{ij} and \mathbf{h}_{rs} is triangulation using least square minimization between these image coordinates to the 3D space (Hartley & Zisserman, 2003; Szeliski, 2010). Finally, we pick a set of neighboring points around the estimated position $\hat{\mathbf{M}}_i$ with the likely location of the true heads, as shown in Fig. 3.2a. These sets of hypotheses laid close to the ground-truth heads, as show in Fig. 3.2b.

3.2. Multi-view Projection

Wide-baseline multiple view configurations are commonly used to cover a scene from a different viewpoints. This configuration allow for spatial reconstruction and the generation of spatial information, similar than ranger sensors or stereo imaging does (Szeliski, 2010). These multiple

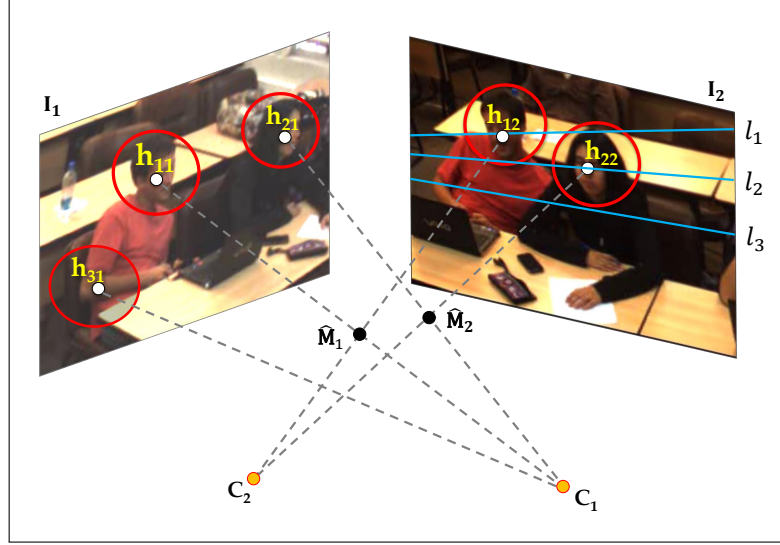


Figure 3.3. Shows the triangulation between head hypotheses from two images I_1 and I_2 at different viewpoints. Hypotheses $\{h_{11}, h_{21}, h_{31}\}$ in image I_1 generate epipolar lines l_1 , l_2 and l_3 in image I_2 . The pairs set of head hypotheses $\{h_{11}, h_{12}\}$ and $\{h_{21}, h_{22}\}$ share the same 3D position \hat{M}_1 and \hat{M}_2 , respectively. We estimate these spatial positions by triangulation using least square minimization along the ray $\overline{h_{ij}C_i}$. Due to there are no head hypothesis near to epipolar line l_3 , the h_{31} do not generate potential head position.

view setups use camera parameters and geometric constraints between viewpoints to execute spatial reconstruction of the environment and the objects within it. Thus, people detection approaches based on wide-baseline multiple view configurations present the advantage to combine and discard false detections using spatial cues. Further, this camera configuration contributes to improve the final detection in situations of heavy occlusion (Mittal & Davis, 2003; Khan & Shah, 2006; Eschel & Moses, 2010; Liem & Gavrilu, 2013). There are geometric techniques based on epipolar geometry and camera calibration that allow us to establish relationships across views in a camera system, providing useful ways to combine and integrate this information (Hartley & Zisserman, 2003; Szeliski, 2010). However, there are practical constraints when appearance of the target object change drastically among viewpoints (Lowe, 2004; Mikolajczyk & Schmid, 2005).

We propose an alternative method to combine detection in wide-baseline configurations in which we generate the corresponding projections of the object located in the 3D space without using standard matching techniques. These projections let us to acquire aspects of people head from various viewpoints simultaneously. Further, using this representation we might include enriched appearance information and 3D cues, *e.g.*, help to locate people in the scene or to filter out regions

where people is not likely to appear. In this step, we compute the sliding-box projections onto the images in order to generate correlated image sections. As other multiple view approaches (Khan & Shah, 2006; Eshel & Moses, 2010; Liem & Gavrilu, 2013), our approach requires a fully calibrated multiple view system of N cameras C_1, \dots, C_N to compute the geometric model. This model relates the 3D world homogeneous coordinates $\mathbf{M}_i = [X_i \ Y_i \ Z_i \ 1]^T$ to the 2D image coordinates $\mathbf{m}_{ij} = [x_{ij} \ y_{ij} \ 1]^T$ in each image I_j . This coordinates relation was obtained for $j = 1, \dots, N$ cameras using the transformation

$$\lambda \mathbf{m}_{ij} = \mathbf{P}_j \mathbf{M}_i, \quad (3.1)$$

where λ is a scale factor and \mathbf{P}_j is the 3×4 calibration matrix of camera C_j (Hartley & Zisserman, 2003).

A *sliding-box* \mathbf{B}_i is a parallelepiped defined in the 3D space as a virtual volume element, which presents the following three properties:

- a) \mathbf{B}_i is centered at 3D point \mathbf{M}_i with coordinates (X_i, Y_i, Z_i) ;
- b) the volume occupied by \mathbf{B}_i belongs to the 3D space of interest S where the object target class will be detected;
- c) it contains a volume of interest with size and shape of \mathbf{B}_i corresponds to the real size and shape of the object target classes.

Follow these properties, the volume of interest of \mathbf{B}_i can be represented using various kinds of shapes to cope with the real size and geometry of the object, *e.g.*, using spheres, ellipsoids, cubes or another parallelepiped. In our research, we are detecting heads. For practical purposes, the *sliding-box* \mathbf{B}_i defines a sphere circumscribed to its occupied space, centered in \mathbf{M}_i and with radius r . This geometric representation fits well with the oval shape of heads. In this detection context, we model mathematically the *sliding-box* as quadric that is a 3D surface defined in 3D world homogeneous coordinates by

$$\mathbf{X}^T [\mathbf{H}^T \mathbf{Q} \mathbf{H}] \mathbf{X} \equiv \mathbf{X}^T \begin{bmatrix} \mathbf{I}_{3 \times 3} & \mathbf{0}_{3 \times 1} \\ \mathbf{0}_{1 \times 3} & -r^2 \end{bmatrix} \mathbf{X} = 0 \quad (3.2)$$

considering

$$\mathbf{Q} = \begin{bmatrix} \mathbf{I}_{3 \times 3} & \mathbf{0}_{3 \times 1} \\ \mathbf{0}_{1 \times 3} & -r^2 \end{bmatrix} \quad (3.3)$$

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_{3 \times 3} & \mathbf{M}_i \\ \mathbf{0}_{1 \times 3} & 1 \end{bmatrix}, \quad (3.4)$$

where \mathbf{Q} defines the quadric shape as a sphere, \mathbf{H} is a transformation matrix which places the volume of interest of \mathbf{B}_i centered at position \mathbf{M}_i , and \mathbf{X} represents the 3D points on the surface defined by the quadric. A surface quadric is projected onto the image I_j as a conic section \mathbf{C} according to the size of the volume element \mathbf{B}_i . This quadric is projected using the camera matrices \mathbf{P}_j as

$$\mathbf{C} = [\mathbf{P}_j \tilde{\mathbf{Q}} \mathbf{P}_j^T]^{-1} \quad (3.5)$$

$$\tilde{\mathbf{Q}} = \mathbf{H}^T \mathbf{Q} \mathbf{H} \quad (3.6)$$

where \mathbf{C} is the conic section defined by the projected quadric on the image I_j , and $\tilde{\mathbf{Q}}$ is a transformed version of \mathbf{Q} after applying \mathbf{H} . Finally, the projection W_{ij} is defined by the subscribed parallelepiped over the conic \mathbf{C} on the image I_j , as shown in Fig. 3.4. More details about quadric and conic representations can be found in Hartley and Zisserman (2003).

We also include a spatial prior or contextual information by defining a space of interest S in 3D as shown in Fig.3.5. This space could be defined as a rectangular parallelepiped where coordinates X , Y and Z are constrained by a lower and a higher bound a and b , respectively *i.e.*,

$$S = \{(X, Y, Z) \mid X \in [X_a, X_b]; Y \in [Y_a, Y_b]; Z \in [Z_a, Z_b]\}. \quad (3.7)$$

We let the sliding-box runs within this space of interest S . This top-down information allows to us limit the action of our sliding-box to areas where we expect to find people or objects in the scene or where we want to search, *e.g.*, we do not expect to find people on the ceiling or laying down on the floor. As is the case in sliding-window approaches, we run the sliding-boxes overlapped to

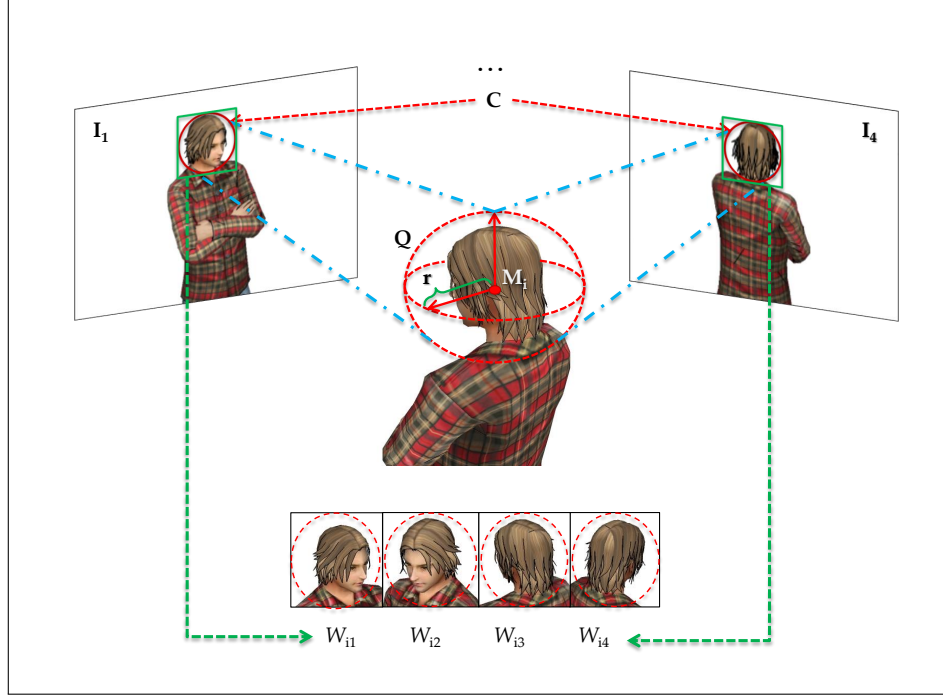


Figure 3.4. Projection diagram of a sphere quadric Q defined on M_i with radius r . In this example, for $N = 4$, Q is projected onto the images I_1, \dots, I_4 as a conic C . The projections W_{ij} are defined as the maximum quadrilateral subscribed over C and showed as dashed red circles in this figure. The elements W_{ij} represent the projection of B_i onto the camera j . All of these elements define a collection of aspects which represents the box B_i seen from each camera. Each element W_{ij} was cropped and then rescaled to 64×64 pixels before feature extraction to cope with projection at different size.

ensure that heads are contained in a whole box. In this way, our method is similar to a sliding-window method, but we can control the space of interest S where our box runs and the 3D size of B_i according to the search requirements.

3.3. Feature Extraction

Feature extraction corresponds to the process in which the visual information in the images pass into a new space of variables less redundant and more informative than the image domain (Szeliski, 2010; Nixon & Aguado, 2012). Advanced detection algorithms use local descriptors to increase the distinction between objects variation or classes, while providing invariance to light changes, blurring, rotation, scaling, noise and differences in viewpoint (Ojala et al., 2000; Viola & Jones, 2001; Lowe, 2004; Dalal & Triggs, 2005; Bay et al., 2008; Leutenegger et al., 2011; Alahi et al., 2012). Most of the features have their own advantages and disadvantages. Recently,

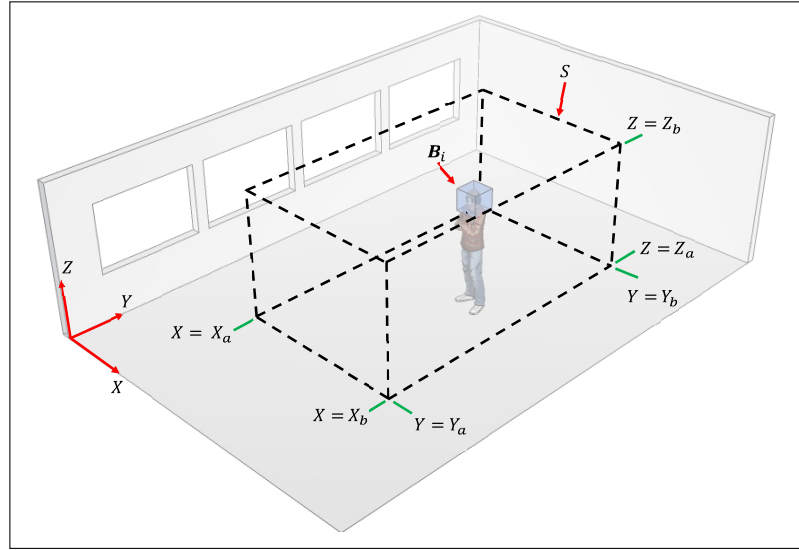


Figure 3.5. Diagram of the space of interest S inside of a room and defined as parallelepiped with set of boundaries $[X_a, X_b]; [Y_a, Y_b]; [Z_a, Z_b]$. We use this contextual information to limit the action of our sliding-box B_i within the space S . This allows to us to search for people's heads in areas in which they are likely to appear according to the context.

region descriptors evaluations agree about that there is no a better descriptors and interest region detectors in all aspects, but their performances depend on the task (Z. Song & Klette, 2013; Wu & Lew, 2013).

In terms of object and people detection, most of the detection algorithms sample the invariant local features from images using an spatial pyramid (Lazebnik et al., 2006; Bosch et al., 2007) or a multi-scale dense grid (Dalal & Triggs, 2005). The last method fits well for multi-scale pedestrian detection when the size of the object is unknown and the algorithm have to uncover them at various scales. In general, Histogram of Oriented Gradient (HOG) (Dalal & Triggs, 2005) and Local Binary Pattern (LBP) Ojala et al. (2000) are state-of-the-art low level features that present similar performance level in people detection. However, it has been reported that LBP improves the detection performance because of this descriptor is invariant to strict monotonic changes in intensity, making it robust against changes in lighting (Wang et al., 2009; Mu et al., 2008).

The LBP is a handcrafted local binary descriptor used to measure local appearance within a neighborhood of P pixels equally spaced on a circle of radius R (Ojala et al., 2000). It calculates

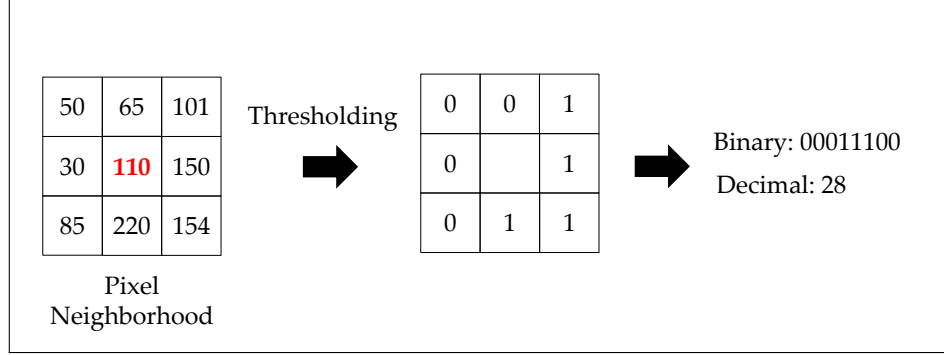


Figure 3.6. Example of the pyramidal feature extraction on a W_{ij} patch. Features are computed in $l = 0, 1, 2$ levels of the image patch. In each level, the image has 4^l cells. The final descriptor has $N_f = 59 \times (1 + 4 + 16) = 1,239$ bins.

the differences between a center pixel with its neighbors, as shown in Fig 3.6. This comparison is represented as a decimal number

$$LPB(x, y) = \sum_{i=0}^{P-1} b_i 2^i, \quad (3.8)$$

where P are the number of neighbors pixels. Ojala et al. (2000) defines two cases of LBP: non-uniform and uniform. A uniform LBP occurs when the binary pattern contains at most two bitwise transitions from 0 to 1 or vice versa over the circular neighborhood, *e.g.*, 00000000 has 0 transitions; 01110000 has 2 transitions. The patterns 11001001 has 4 transitions and 01010010 has 6 transitions. Both are non-uniform patterns. In the uniform case, each uniform LBP pattern is cast into a unique histogram bin according to its decimal value. All non-uniform LBP patterns are cast in single bin. In this way, an 8 element neighborhood has a total of 256 patterns, 58 of which are uniform, and therefore, the final descriptor is an histogram with 59 bins. In general, 3x3 pixel neighborhood over the uniform quantization achieves best performances for most applications (Wang et al., 2009; Mu et al., 2008; Vedaldi & Fulkerson, 2010). We defer the reader to Ojala et al. (2000) and Mu et al. (2008), for a detailed explanation of the LBP descriptor.

We represent each window W_{ij} as a bounding-box in I_j defined as the maximum subscribed parallelepiped in the projection of box \mathbf{B}_i as a conic \mathbf{C} in accordance with (3.5). Each projection W_{ij} was rescaled to 64×64 pixels to allow for the different sizes of the projections. Multi-scale pyramid decomposition has been the standard sampling strategy for detecting objects at unknown scales in images (Dalal & Triggs, 2005; Felzenszwalb et al., 2010). However, we project known size



Figure 3.7. Example of the pyramidal feature extraction on a W_{ij} patch. Features are computed in $l = 0, 1, 2$ levels of the image patch. In each level, the image has 4^l cells. The final descriptor has $N_f = 59 \times (1 + 4 + 16) = 1,239$ bins.

objects originating in projections at different scales simultaneously in each view. After resizing, we extract a set of features in pyramidal decomposition for each window W_{ij} (Lazebnik et al., 2006; Bosch et al., 2007), where each W_{ij} is represented by a feature vector \mathbf{x}_{ij} . This decomposition allows to us to extract global and local information from each instance. Each level $l \in L = \{0, \dots, n\}$ in the pyramid has 4^l cells or patches, and for each cell we compute a descriptor with K bins. The descriptor of the entire image patch W_{ij} has $N_f = K \sum_{l=0}^L 4^l$ bins, as shown in Fig. 3.7.

3.4. Multiple View Classifier

The classification model is an important part for all machine learning method. This model evaluates whether a set of features belongs to a specific class. Most of detectors based on sliding window use a discriminative approach due to their high level of performance, embedded ability to select relevant features, and ease of use. The most popular discriminative models are Support Vector Machines Cortes and Vapnik (1995) and Boosting Schapire (1990). One of the advantages that captures the attention on SVMs is their ability to build non-linear decision boundaries using the a kernel trick (Papageorgiou & Poggio, 2000; Dalal et al., 2006; Wang et al., 2009; Schwartz et al., 2009; Felzenszwalb et al., 2009; Pedersoli et al., 2014). Moreover, SVMs have faster training speed. However, the runtime complexity of a non linear SVM classifier is high. Boosted decision trees on the other hand have faster classification speed but are significantly slower to train and the complexity of training can grow exponentially with the number of classes (Torralba et al., 2004). Due to this facts and plus the mathematical background, linear kernel SVMs have become popular

for real-time applications as they enjoy both faster training and classification speeds (Maji et al., 2008; Yang et al., 2009).

Once we extract the features for each element W_{ij} , we apply one classifier in order to identify the collection of aspects or the multiple view projections simultaneously. We apply two different strategies to classify the multiple view projections: ensemble of features and ensemble of classifiers. In the experiments, we evaluate both schemes independently in order to present pros and cons of them. The feature ensemble scheme describes a sliding-box using one descriptor that is built by concatenating the features present in each view individually. A single classifier is used to identify these compounded features. The ensemble of classifiers applies a set of classifiers to each view in order to generate a set of mid-level features. These new features describe a sliding-box and a classifier learns these new mid-level features. The next two sub-sections describe how we trained both classification approaches for projection sequences. As we mentioned previously, both are independent and exclusive of each other.

3.4.1. Ensemble of Features

The ensemble of features consist in combining features as a single descriptor where each feature contributes to discriminate the object target class. For example, Wang et al. (2009) propose an algorithm that uses as input a combination of HOG and LBP descriptors. This combination helps to discriminate whether a detected windows is occluded. In each descriptor contributes with an specific property of the object. In 2014, Pei et al. prose a pedestrian detector based on combinations of HOG and LBP features. Experiments include combining low-level features and transformed versions of these features using Principal Component Analysis (PAC) and Singular Value Decomposition as K-SVD. Results show that combination outperform original version of HOG and LBP and helps to filter false positive detections.

One of the advantages of our approach is the ability to retrieve the multiple views projections of a head simultaneously. This advantage allows us to generate enriched feature descriptors by combining the descriptors coming from each view. This new descriptor \mathbf{x}_i represent an enriched appearance of head. In this scenario, we train a single SVM classifier as described in Chapter 2.2.1, which scores an example \mathbf{x}_i with a function of the form

$$f_{\beta}(\mathbf{x}_i) = \beta \cdot \Phi(\mathbf{x}_i), \quad (3.9)$$

where β is the vector of model parameters and Φ is a kernel function to transform the example \mathbf{x}_i . We find the model parameters from a labeled dataset $D = (\langle \mathbf{x}_1, y_1 \rangle, \dots, \langle \mathbf{x}_n, y_n \rangle)$, by minimizing the objective function

$$L(\beta) = \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \max(0, 1 - y_i f_{\beta}(\mathbf{x}_i)) \quad (3.10)$$

where C is a cost constant for increasing the penalization of the classification error; $\max(0, 1 - y_i f_{\beta}(\mathbf{x}_i))$ is the standard hinge loss function; and \mathbf{x}_i is the example instance represented as the feature vector. This is equivalent to the soft margin representation explained in Chapter 2.2.1.

The instance $\mathbf{x}_i = [\mathbf{x}_{i1}, \dots, \mathbf{x}_{iN}]$ represents a feature vector formed by concatenation of single descriptors in order to simultaneously evaluate the collection of aspects, as shown in Fig. 3.8. Each W_{ij} is associated with a feature \mathbf{x}_{ij} vector with N_f bins. If the camera system has N cameras, this descriptor has $N_s = N \times N_f$ bins. The key idea is to build up an enriched feature vector which represents the head structure in a global perspective as shown in Fig. 3.4. This training process yields a model β_{fe} , which we use to evaluate each box along the detection process. This model represents the head or any object structure using this enriched feature vector taking into account various viewpoints simultaneously.

3.4.2. Ensemble of Classifiers

The idea of ensemble methods consist in using multiple classification models strategically generated and combined to solve a particular learning problem (Polikar, 2006). There are well know examples of successful classifiers ensembles such as Boosting (Freund & Schapire, 1995), Bagging (Breiman, 1996) and Random Forest (Breiman, 2001) that demonstrate the improvements in prediction performance. Ensembles also allows for selecting optimal features sets, data fusion, incremental learning and error correcting.

As we mention before, our approach allows us to retrieve the multiple views projections of a head simultaneously. Instead of using the descriptors directly into the classifier, we can include a classification stage that helps to reduce the variability of the features. In this approach, we build

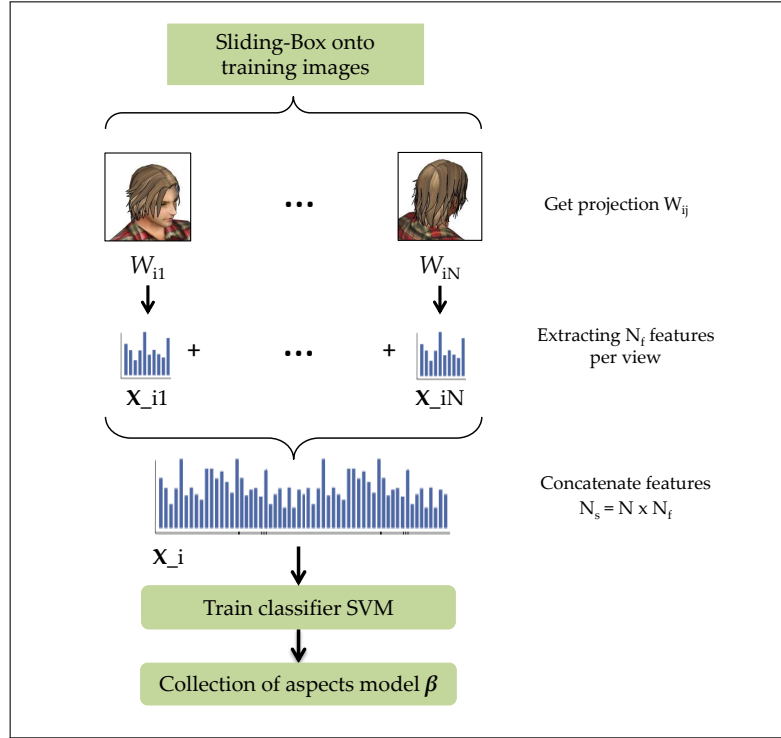


Figure 3.8. Training process diagram of features ensemble scenario. Once we extract features from each element W_{ij} , we concatenate all of the N_f features in a single descriptor with N_s bins. We use a single SVM classifier to learn a model β_{fe} using examples of head sequences.

an ensemble of classifiers composed of two layers, where the first layer contains a set of mid-level features which summarize the feature vectors \mathbf{x}_i ; and the second layer ensembles the mid-level features to classify the collection of aspects W_{ij} represented by the instance \mathbf{x}_{ij} .

The first layer is a linear multi-class SVM model which learns to discriminate among k classes and assigns a confidence score $f_{\beta}^k(\mathbf{x}_i)$ to each class, where these classes represent head aspects and background. This layer of classifiers transform a set of feature vectors \mathbf{x}_{ij} into a set of mid-level features $[\{f_{\beta}^1(\mathbf{x}_{i1}), \dots, f_{\beta}^k(\mathbf{x}_{i1})\}, \dots, \{f_{\beta}^1(\mathbf{x}_{ij}), \dots, f_{\beta}^k(\mathbf{x}_{ij})\}, \dots, \{f_{\beta}^1(\mathbf{x}_{iN}), \dots, f_{\beta}^k(\mathbf{x}_{iN})\}]$, where N is the number of cameras in vision system. Opposite to classification problems which require $\max_{j=1, \dots, k} (\beta^j \Phi(\mathbf{x}_i))$ to discriminate among classes, we keep all decision values product to apply each model k as features for the second layer.

We address the SVM multi-class problem using two methods: one-against-all and one-against-one. Both methods of multi-class classifiers were trained using a bootstrap strategy, which is described in next section. The one-against-all constructs k SVM models where the i -th model is trained using the examples in the i -th class as positive instances, and all of the other examples in other poses as negative instances. After solving (3.10) for each class there will be k decision functions $\langle f_{\beta}^1(\mathbf{x}_i) = \beta^1 \cdot \Phi(\mathbf{x}_i) \rangle, \dots, \langle f_{\beta}^k(\mathbf{x}_i) = \beta^k \cdot \Phi(\mathbf{x}_i) \rangle$. The one-against-one method constructs $k(k-1)/2$ classifiers trained on data from two classes, where the classification problem is defined as:

$$\min_{\beta} \frac{1}{2} \|\beta^{s,t}\|^2 + C \sum_{i=1}^n \max(0, 1 - y_i f_{\beta}^{s,t}(\mathbf{x}_i)) \quad (3.11)$$

As we had done with the one-vs-all method, we only use the decision values to represent the collection of aspects by a collection of scores.

In the second layer, a new SVM model learns from the set of scores $f_{\beta}^1(\mathbf{x}_{ij}) \dots f_{\beta}^k(\mathbf{x}_{ij})$ built in the previous layer to identify the entire collection of aspects projected from the sliding-box \mathbf{B}_i . This final model β can merge the information coming from the camera system in the detection stage. The feature vector at the second layer has $N_s = N_{scores} \times N$ elements, as shown in Fig. 3.9.

3.4.3. Bootstrap Training

We train both systems using a bootstrap strategy to avoid biasing the training set, memory overloads, and over-fitting the model (Felzenszwalb et al., 2009). A bootstrap algorithm is an iterative procedure which modifies the train dataset along the training rounds by including a ratio of misclassified samples and releasing a ratio of properly classified ones. Let $D = \{\langle \mathbf{x}_1, y_1 \rangle \dots \langle \mathbf{x}_n, y_n \rangle\}$ be a fully labeled dataset with features \mathbf{x}_i and labels $y_i \in \{-1, 1\}$. The training procedure starts with a random subset $X \subset D$ used to train a model β . Then we apply this model to the entire dataset D in order to distinguish between hard examples H and easy examples E associated with β . The algorithm considers an example \mathbf{x}_i hard if $f_{\beta}(\mathbf{x}_i) > 1$. On the other hand, an example is considered to be easy if $f_{\beta}(\mathbf{x}_i) < 1$. We randomly select a sample $H_s \subset H$ and $E_s \subset E$ according to the decision value f_{β} of each \mathbf{x}_i . The selected hard examples H_s will be added to X , and the selected easy examples E_s will be removed from X . This modified version of X is used in the next training round. The algorithm iterates a fix number of times. Finally, this procedure generates the model

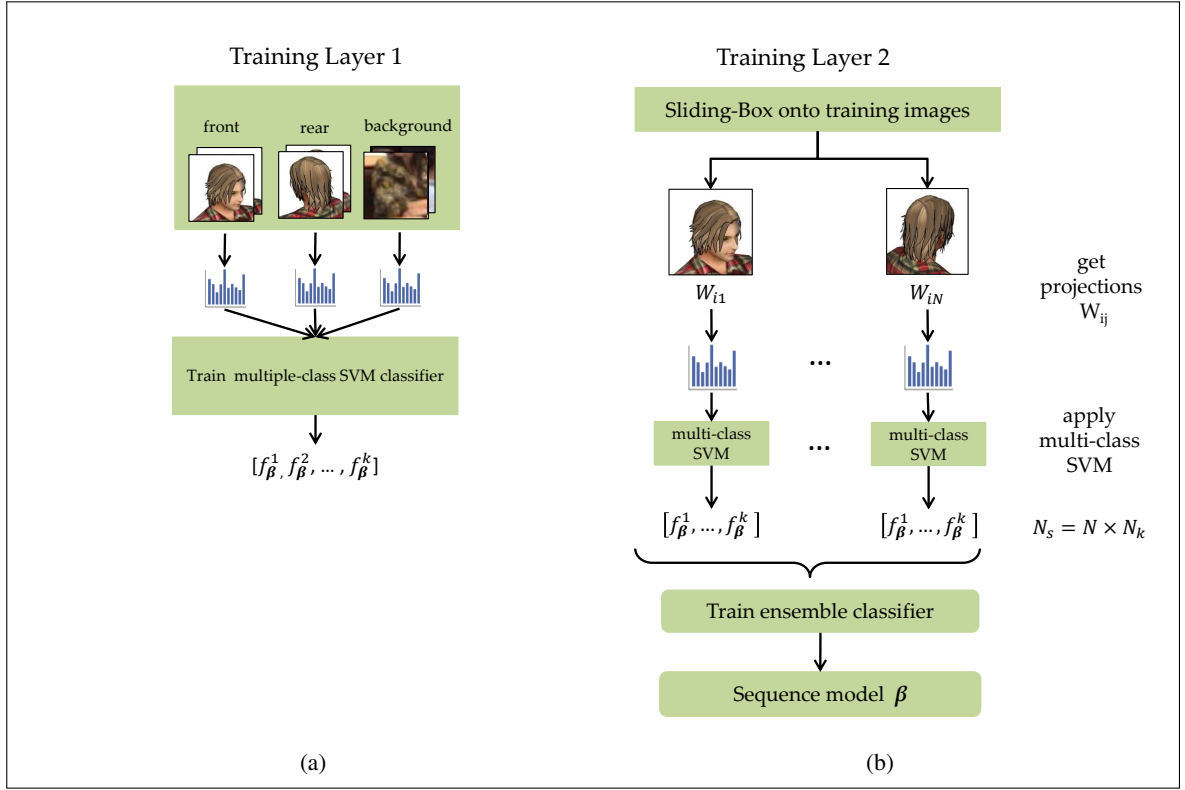


Figure 3.9. Training process diagram for the classifier ensemble scheme. We use an ensemble of classifiers divided into two layers: (a) shows the first layer, which is formed by multi-class SVM classifiers. This layer can identify frontal head, rear head, and background. (b) shows the second layer, which is trained using the scores $(f_{\beta}^1, \dots, f_{\beta}^k)$ obtained by applying the first layer of classifiers to each element W_{ij} . This process yielded the model β , which can classify the image sequence.

used to classify the collection of aspects during the detection process. Algorithm 1 shows all of the steps of the training process.

3.4.4. Best Collection of Aspects

Classifiers described in Sections 3.4.1 and 3.4.2 discriminate over aligned collection of aspects that presents a sequence of aspects for a global head pose. However, during detection we do not have a priori knowledge about the head aspect in the scene. To face with this uncertainty, we generate four collection of aspects from the input collection applying circular shifts on it. These shifts allows us to find the most confident collection of aspects in accordance with the template used for training. We apply the multiple view classifier to each collection and we choose the one with highest confidence using an *argmax* criterion, as shown Fig. 3.10. We do not use random

Algorithm 1: Bootstrap train

```
Let  $D = \{\langle x_1, y_1 \rangle \cdots \langle x_n, y_n \rangle\}$ ,  $y_i \in \{-1, 1\}$  a full labeled dataset
Let  $X \subset D$  a random set
while ( $\text{iter} \leq \text{max-iter}$ ) do
   $\beta \leftarrow \text{train\_model\_svm}(X)$ 
   $f_\beta \leftarrow \text{apply\_svm\_model}(\beta, D)$ 
   $H(\beta, D) \leftarrow \{\langle x_i, y_i \rangle \in D \mid (f_\beta(\mathbf{x}_i) < 1)\}$ 
   $E(\beta, D) \leftarrow \{\langle x_i, y_i \rangle \in D \mid (f_\beta(\mathbf{x}_i) > 1)\}$ 
   $H_s \leftarrow \text{select\_random}(H)$ 
   $E_s \leftarrow \text{select\_random}(E)$ 
  addhard( $H_s$ ) to  $X$ 
  remove\_easy( $E_s$ ) from  $X$ 
end while
return svm\_model  $\beta$ 
```

shifts or random combination of aspects, because an admissible collection of head aspects present a coherent sequence of appearance according to our camera system, *e.g.*, the collection of aspects *rear* – *front* – *rear* – *front* is not allowed.

3.5. Non Maximal Suppression

As we discussed in Chapter 2.4, a Non-Maximal Suppression (NMS) is a critical procedure in computer vision algorithms in which one must choose the most representative detection from a set of confident multiple detections. In general, the most basic approach consists in processing all detection windows using an agglomerative clustering algorithm in order to filter the spurious responses and to merge those that are overlapping. Dalal (2006) propose a clustering based suppression procedure that locates the local modes from a set of multiple detections using a weighted version of the Mean Shift algorithm (Fukunaga & Hostetler, 1975; Y. Cheng, 1995; Comaniciu & Meer, 2002). The algorithm considers that each detection $\mathbf{y}_i = (x_i, y_i)$ has an associated symmetric positive definite 3×3 bandwidth covariance matrix \mathbf{H}_i to define a smoothing kernel for the detected position (x_i, y_i) . Finally, the algorithm estimates the density of overlapped detections to represent these n points as local modes. In 2009, Dollar et al. propose a Local Maximum Searching (LMS) as suppression procedure. Felzenszwalb et al. (2009) use the same NMS approach to suppress multiple detections. The LMS algorithm chooses a final detection by applying a greedy strategy that suppresses the less confident of every pair of detection sufficiently overlapped. This

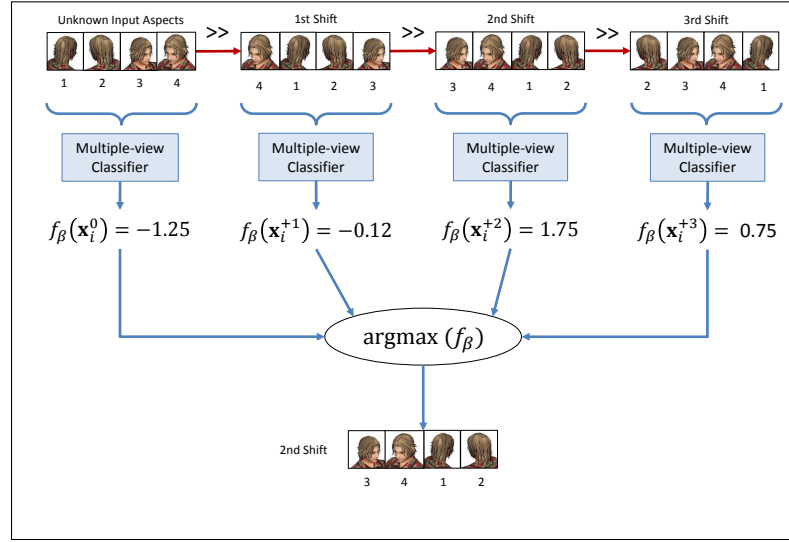


Figure 3.10. Diagram of best collection searching. The algorithm receives an input collection of aspects without a priori knowledge about its correct alignment. We apply a set of circular shifts to generate the total number of four collections of aspects, including the input collection. After applying the multiple view classifier to each collection, we choose the most confident one using an *argmax* criterion. In the diagram, the multiple view classifier assigns a set of confidence values to each collection of aspects: $f_{\beta}(x_i^0) = -1.25$, $f_{\beta}(x_i^{+1}) = -0.12$, $f_{\beta}(x_i^{+2}) = 1.75$, $f_{\beta}(x_i^{+3}) = 0.75$. Finally, our algorithm selects the collection of aspects generated with the second shift because it best matches the training samples.

procedure assigns the detection y_i the maximum score around its neighborhood. The overlap criteria is defined in terms of the Euclidian distance between detections.

There are recent algorithms that attempt to improve suppression procedures based only in clustering techniques. In 2012, Zaytseva and Vitria propose a search based approach to non-maximum using a statistical framework. The algorithm requires a discriminative model previously trained to generate a prior distribution. Finally, a Markov chain Monte Carlo performs. Results show a promising use for the prior distribution in to search and merge potential detection. However, final detections depends on the discriminative model that may miss true positive detection affecting the overall performance. Shuai et al. (2012) propose a hierarchical clustering based NMS method applied to pedestrian detection. This algorithm uses a clustering methods based on ellipse Euclidean distance to get the location of pedestrian. The non-maximum suppression process is partitioned into two parts hierarchically. Results show that the non-hierarchical clustering based method perform

similar to the weighted Mean Shift proposed in (Dalal, 2006), but consumed less time. On the other hand, the hierarchical algorithm recalled more true positives than the non-hierarchical method.

Due to the stability and the use on multiple datasets, we prefer to implement a modified version of the Weighted Mean Shift (WMS) NMS (Dalal, 2006) and the Local Maximum Searching (LMS) (Dollar et al., 2009) to work in the 3D domain. Both procedures require a set of sliding-boxes \mathbf{B}_i , $i = 1 \dots n$ provided by the our detector, where each detection is defined by its 3D location and size. Then, the NMS finds the most confident box within a neighborhood.

Our implementation of the LMS algorithm chooses a final detection by applying a strategy in (Dollar et al., 2009) to suppresses the less confident sliding-box of every pair of sliding-boxes \mathbf{B}_i sufficiently overlapped. Our procedure assigns the sliding-box \mathbf{B}_i the maximum score f_β . We define the overlap criterion in terms of the Euclidian distance between the sliding-boxes. This procedure only requires a distance threshold as a parameter because all of the sliding-boxes have the same size and shape.

The WMS algorithm is used to locate the local modes from a set of multiple detections (Dalal, 2006). Each detection \mathbf{B}_i has an associated symmetric positive definite 3×3 bandwidth covariance matrix \mathbf{H}_i to define the smoothing width for the detected position \mathbf{M}_i . Overlapped detections are fused to represent the n points as local modes. The derivation is the same as in (Dalal, 2006). We assumed the smoothing kernel as Gaussian, therefore the weighted kernel density estimate at a point is given by

$$\hat{f}(\mathbf{M}) = \frac{1}{n(2\pi)^{3/2}} \sum_{i=1}^n |\mathbf{H}_i|^{-1/2} t(\varpi_i) \exp \left(-\frac{D^2[\mathbf{M}, \mathbf{M}_i, \mathbf{H}_i]}{2} \right) \quad (3.12)$$

where

$$D^2[\mathbf{M}, \mathbf{M}_i, \mathbf{H}_i] \equiv (\mathbf{M} - \mathbf{M}_i)^T \mathbf{H}_i^{-1} (\mathbf{M} - \mathbf{M}_i) \quad (3.13)$$

is the Mahalanobis distance between \mathbf{M} and \mathbf{M}_i , and the term $t(\varpi_i)$ provides the weights for each detection assigned by the classifier. These weights are defined as

$$\varpi_i(\mathbf{M}) = \frac{|\mathbf{H}_i|^{-1/2} t(\varpi_i) \exp \left(-D^2[\mathbf{M}, \mathbf{M}_i, \mathbf{H}_i]/2 \right)}{\sum_{i=1}^n |\mathbf{H}_i|^{-1/2} t(\varpi_i) \exp \left(-D^2[\mathbf{M}, \mathbf{M}_i, \mathbf{H}_i]/2 \right)} \quad (3.14)$$

The mode is reached by iteratively computing

$$\mathbf{M}_m = \mathbf{H}_h(\mathbf{M}_m) \left[\sum_{i=1}^n \varpi_i(\mathbf{M}_i) \mathbf{H}_i^{-1} \mathbf{M}_i \right] \quad (3.15)$$

for each detection at location \mathbf{M}_i until it converges to the local mode \mathbf{M}_m , *i.e.*, the new mode location no longer changes. The term \mathbf{H}_g is the weighted harmonic mean of the covariance matrices \mathbf{H}_i . The set of all modes represents the final detections where the modes are the detection locations and the mode peaks are the final scores. For each point, the algorithm is guaranteed to converge with the mode. For more detailed information on the weighted Mean Shift algorithm and its properties, see (Comaniciu & Meer, 2002; Dalal, 2006).

We assume diagonal covariance matrices \mathbf{H}_i only with the uncertainty of location because our sliding-box has the same size and shape. Let $\text{diag}[\mathbf{H}_i]$ represent the 3 diagonal elements of \mathbf{H}_i , such that

$$\text{diag}[\mathbf{H}_i] = [\sigma_x^2, \sigma_y^2, \sigma_z^2] \quad (3.16)$$

where $\sigma_x, \sigma_y, \sigma_z$ are the user supplied smoothing values.

We utilize the suggested transformation function for SVM, which uses a hard clipping to ensure positive weights while running the NMS procedure (Dalal, 2006), such that

$$t(\varpi) = \begin{cases} 0 & \text{if } \varpi < c \\ w - c & \text{if } \varpi \geq c \end{cases} \quad (3.17)$$

where c is a threshold which controls the weight values and helps avoid false detections to get into the Mean Shift algorithm.

Chapter 4. METHODOLOGY AND IMPLEMENTATION DETAILS

This chapter describes the methodology that we used to demonstrate our hypothesis by applying our approach to real scenes. This methodology includes building a testing environment and establishing a test protocol of our framework. During experimentation we presented our advances as concept tests and preliminary findings in three articles. In order to carry out the people detection experiments, we used a classroom to emulate a controlled indoor environment with people inside. Our experimental methodology includes building training and testing datasets, measuring detection performance, and measuring the improvements that can be made by using complementary data generated by multiple views. Details about hardware, implementation details, and performance measurements will be described in the next three sections.

4.1. Hardware

For experimentation purposes, we mounted a multiple view indoor environment in a classroom. This environment consists in to a set of four Point Grey Flea2 model cameras. All of them were calibrated and synchronized to ensure geometric matching, and the same temporal information simultaneously. Thus, our multiple view camera system has $N = 4$ calibrated cameras. These cameras were installed in the upper corners of the classroom, as shown in Fig 4.1. We used a standard chess-board method for calibrating the cameras included in the calibration toolbox developed for the OpenCV library and MATLAB, which works using a calibration model inspired by Heikkila and Silven (1997) and Z. Zhang (1999). We obtained an average calibration error of ~ 0.4 pixels per camera.

4.2. Datasets Details

We used our camera system in the indoor environment to build our own multiple view head dataset for training and testing. All images were acquired at 640×480 pixels and 15fps. We designed an interactive user interface (GUI) as shown in Fig.4.2, where the user enclosed each ground-truth detection in each camera. We establish their 3D re-projection geometrically according to the geometric model defined previously in (3.1), where a 3D position \mathbf{M}_i is mapped into a 2D position \mathbf{m}_{ij} through the calibration matrix \mathbf{P}_j .

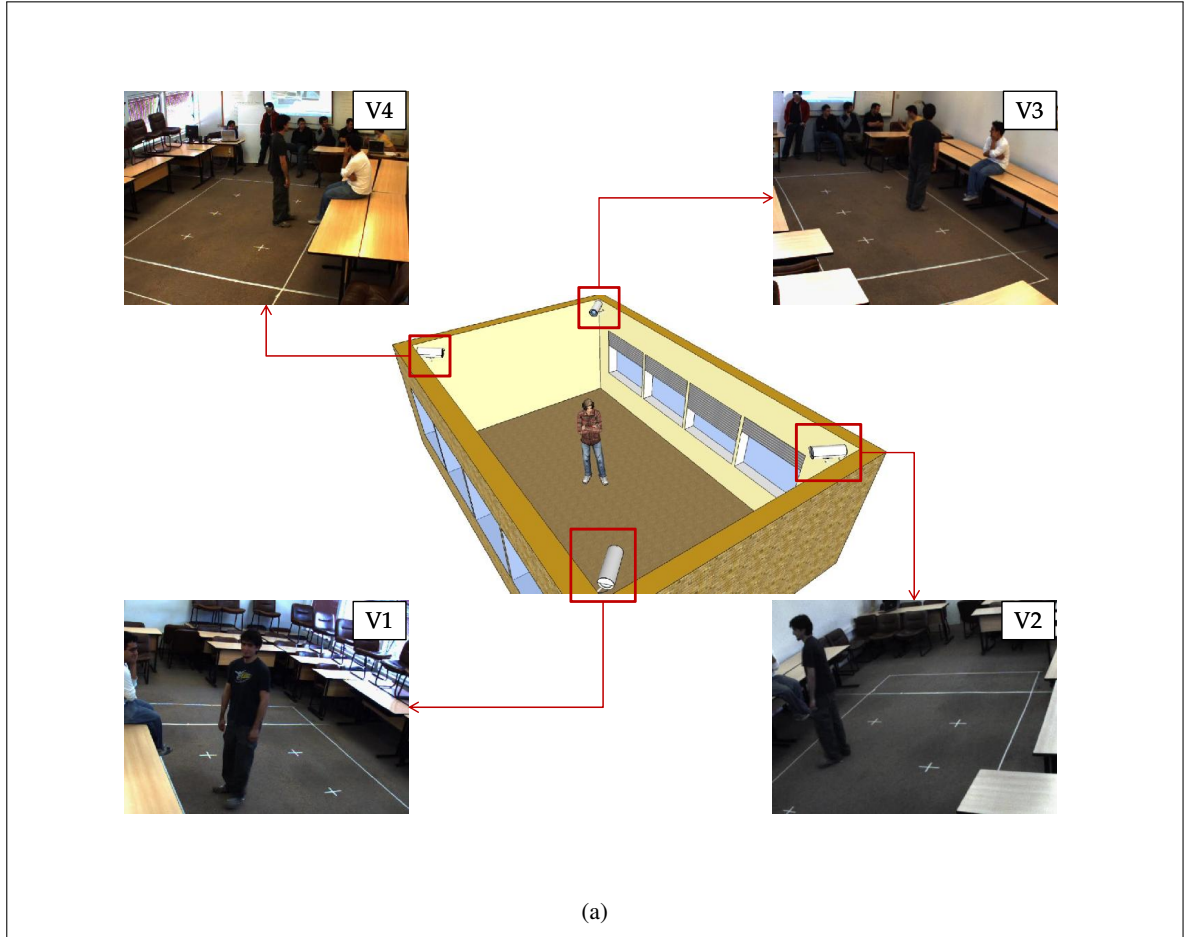


Figure 4.1. We mounted four Point Grey Flea2 cameras in a classroom to simulate an indoor environment where we could detect people. The four cameras were calibrated to ensure the geometric reconstruction of 3D points and synchronized to acquire the images at same time.

4.2.1. Train Dataset

The train datasets contain images in which a set of ten people were placed at six locations within a classroom, as shown in Fig. 4.3. We selected images in which people appear in all of the cameras simultaneously. People spin over their Z axis from 0° to 360° . We manually marked and labeled all positive instances.

The features ensemble method requires samples of collection of aspects that have a predetermined order. This collection of aspects are sets of the sliding-box projections from various view-points. We manually selected the aspects that were most consistent with the method described in

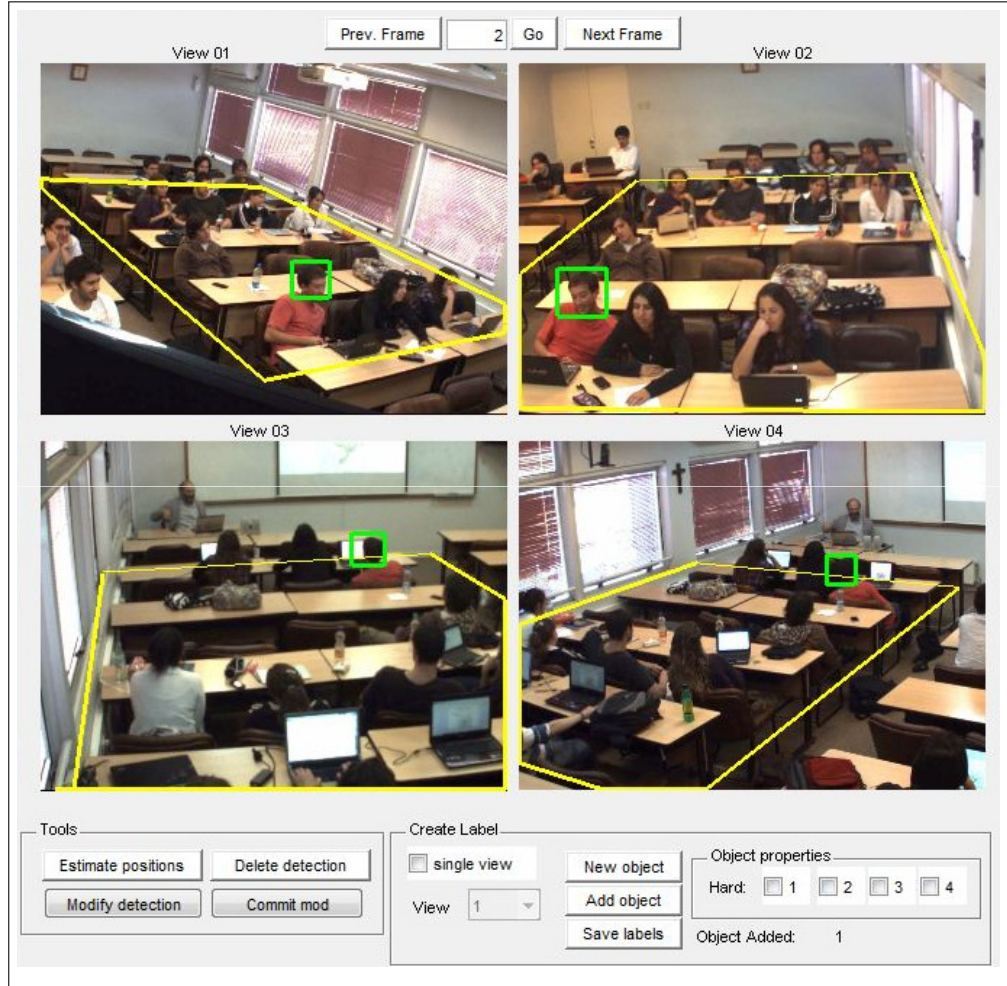


Figure 4.2. Example of our GUI for labeling a set of four images representing the same scene at different viewpoints. A head is labeled selecting two matching points and we estimate its 3D location throughout the geometric model. This location is re-projected onto the images as bounding boxes which represent the head in all viewpoints.

Section 3.4. We also used mirrored versions of train instances to generate new views and enhance the train dataset. Positive instances are the sliding-boxes projected over the people’s location in order to obtain their four projections W_{ij} . We are thus able to generate image aspects for people’s heads as shown in Fig. 4.4. Negative instances are sliding-box projections from random locations in the room that exclude head positions, *i.e.*, clothes, carpets, walls, etc. We also combined individual samples randomly in order to build artificial negative sequences and enrich the dataset. Table shows 4.1 a summary of the train dataset used for the feature ensemble method.

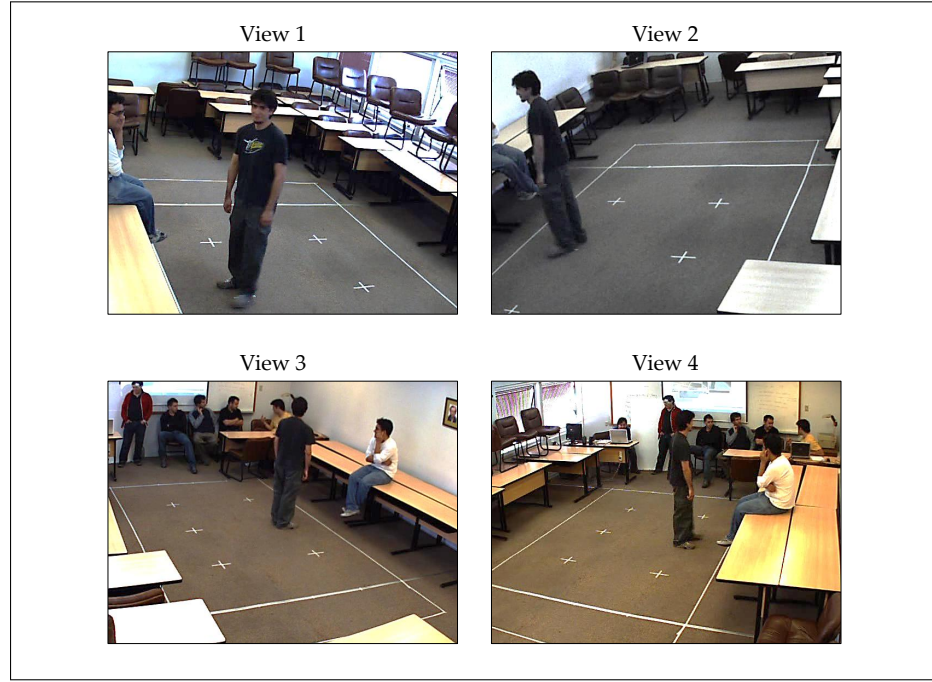


Figure 4.3. Example of images collected for training. Each image comes from one view in the camera system. People stand over the white crosses on the floor and spin around their Z axis to generate various views of the head.

Table 4.1. Details about the training dataset used to train the feature ensemble. Each instance is a collection of aspects, as shown Fig. 4.4

class	number of examples
head (positives)	1,000
background (negatives)	4,570
total	5,570

The ensemble of classifiers uses two training stages. In the first stage, a multi-class classifier learns based on individual aspects separated into three classes: frontal head, rear head, and background. We collected independent instances W_{ij} from the dataset used to train the ensemble of features. This individual dataset contains 16,494 examples separated into three mentioned classes as shown in Table 4.2. In the second stage, an ensemble of classifiers uses the outputs obtained after applying first layer classifiers. We used the same instances described in Table 4.1 to train this ensemble. Although we consider two individual aspects, front and rear, our approach is not limited to the number of poses for individual aspects and collection of aspects.

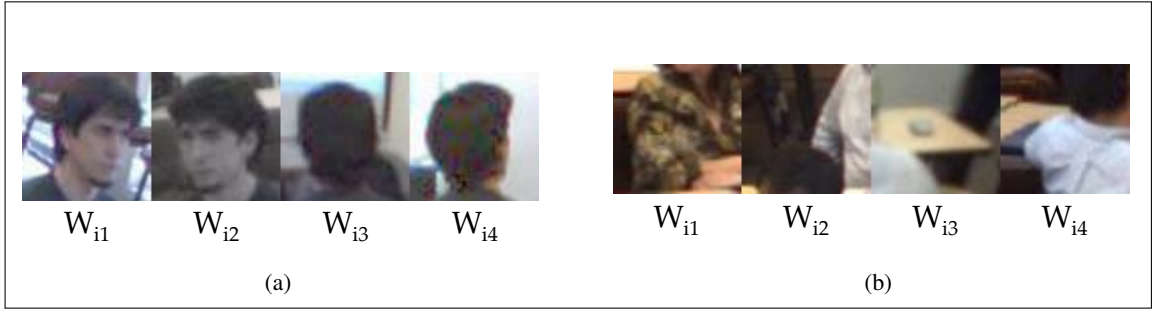


Figure 4.4. Examples of aspects collection used for training. (a) positive instance of people's head retrieved from multi-view camera system and (b) background samples of classroom environment. We used four cameras in both examples, where $j = 1, \dots, 4$

Table 4.2. Details about the training dataset used for training individual models.

class	number of examples
Frontal head	2,000
Rear head	2,000
Background	12,494
total	16,494

4.2.2. Test Dataset

The test dataset consists of two fully labeled multi-view video sequences acquired in a classroom or auditorium environment in which individuals present different activity levels. Both testing sequences contain views of heads and torsos. We labeled an average number of 26,000 instances for testing, taking into account labels in each camera. However, we tested our algorithm over an average of 6,500 detection instances. An instance is defined as a detection which appears in all cameras simultaneously. Table 4.3 shows a summary about both test sequences.

The first sequence, *sq-01*, contains 245 frames of ten people inside of the classroom moving and changing their poses, as shown Fig. 4.5a. This sequence contains 6,800 labels and 1,700 detection instances in average. The sequence presents the main challenge of changing poses and displacements. The second sequence, *sq-02*, contains 600 frames of an average of ten people who were sitting and following a speaker. The subjects were different from those used in the train dataset, as shown Fig. 4.5b. This sequence contains 20,000 labels and 5,000 detection instances in average. The sequence presents the main challenge of individuals who are not seen and a larger set of views.

Table 4.3. Details about test dataset used for testing the detector. Sequences were called sq-01 and sq-02. We show an average number of people per image and the total number of frames in each sequence.

sequence	avg. no. people/image	no. of frames	avg. no. labels	avg. no. ground-truth instances
sq-01	7	245	6,000	1,700
sq-02	8	600	20,000	5,000

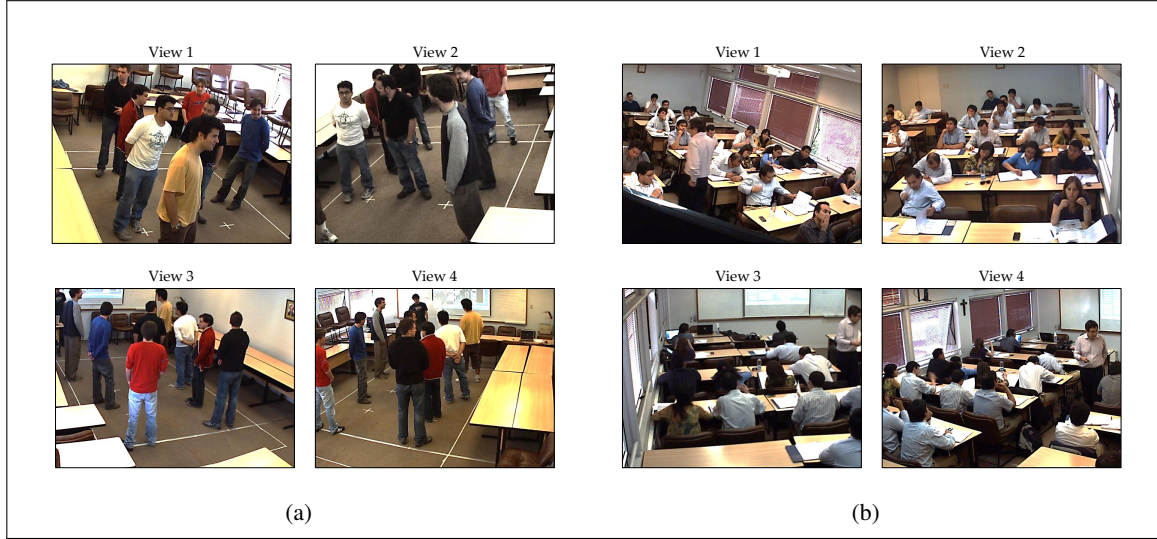


Figure 4.5. Example of images collected for testing. Both test datasets contain images of people in a classroom at different activity levels and under various occlusion conditions. (a) The first test sequence contains people moving and changing their appearances. (b) The second test sequence consists of people sitting observing a lecture. Their appearances change less than the first sequence, and although there is occlusion, it is less frequent.

4.3. Evaluation Methodology

We evaluated our method quantitatively and qualitatively. Precision-recall curves and performance indicators represent quantitative experiments. We used visual observation to perform qualitative evaluations during the experiments. We run tests to measure the ability of the classifier to learn the information shared by the views and how this impacts the separability between classes. We also measure the ability of the method to detect people in real images in an indoor environment. These last experiments are based on a comparison of four methods. The first two correspond to a mono-focal-based method and the same mono-focal method using epipolar geometry to achieve the integration of the detections across multiple views. The last two correspond to our approach following two methodologies: ensemble of features and ensemble of classifiers.

4.3.1. Detection Evaluation

In order to quantify the performance of the detectors, we evaluated all methods proposed in this thesis following the procedure laid out in the PASCAL challenge by Precision-Recall (PR) curves (Everingham et al., 2010). Most of the state-of-art detectors participate in this challenge, and compare performances levels using this standard evaluation. In this context, the precision-recall curve is computed from each methods ranked output, where recall is defined as the proportion of all positive examples ranked above a given rank, and precision is the proportion of all examples above that rank which are from the positive class. Precision-Recall curves close to recall 1 and precision 1 represent better performance. The entire shape of the curve is represented by the average precision-recall (AP) as a single reference value. The computer vision community accept the use of precision-recall curves and the AP value over area under curve measure of the ROC curve because of improvements in the sensitivity of the metric, improvements in interpretability especially for image retrieval applications, and increased visibility to performance at low recall (Everingham et al., 2010).

The AP defined as the arithmetic mean computed on the interpolated precision values $\tilde{p}(rc)$ for 11 thresholds on recall $rc \in 0, 0.1, \dots, 0.9, 1$ such that,

$$AP = \frac{1}{11} \sum_{rc} \tilde{p}(rc). \quad (4.1)$$

The interpolated precision $\tilde{p}(rc)$ represents the maximum precision for which the corresponding recall is greater than or equal to the threshold rc (Everingham et al., 2010). Detection algorithms yield a list of bounding boxes (BB) and a score for each detection. The scores denote a confidence level for each detection, where larger values indicate higher confidence that the object is present.

In order to avoid multiple detections, almost all detection algorithms include a NMS procedure for merging nearby detections. The PASCAL measure evaluates a potential match between detection (BB_{dt}) and ground truth (BB_{gt}) using the area of overlap (a_o) defined as

$$a_o \doteq \frac{area(BB_{dt} \cap BB_{gt})}{area(BB_{dt} \cup BB_{gt})} > 0.5 \quad , \quad (4.2)$$

whose value must exceed 50%, which is the overlap value proposed in the PASCAL Challenge. Detections with a greater confidence level are at first matched greedily with their ground-truth, and the other unmatched BB_{dt} count as false positives. Therefore, multiple overlapping detections are penalized.

We used detection experiments to compare the performance of four methods: 2D deformable part-based model (DPM) detector (Felzenszwalb et al., 2010); 2D deformable part-based model with multiple view filtering (DPM-epipolar); 3D sliding-box with ensemble of features classifier (EF); 3D sliding-box using the one-vs-all ensemble of classifiers (OVA-EC); and the 3D sliding-box using the one-vs-one ensemble of classifiers (OVO-EC). The experiments are performed using the test datasets described in Section 4.2.

4.3.2. Heat Maps

To visualize and to corroborate the spatial behavior of the detections, we used heat maps to identify the most frequently visited areas. These maps contains the cumulative sum of detections, and they are built using the projection of the detections over the plane Z , similar to the occupation maps proposed in Fleuret et al. (2008). Thus, we can observe the behavior of the detections during the execution of the algorithm. Each detection is accumulated over a neighborhood around its location. The more occurrences of detections over a location in the Z plane, the greater the intensity on the map. The final map is displayed in pseudo color, as shown Fig. 4.6.

4.3.3. Per-Windows Evaluation

During this evaluation, we proposed training a classifier using a small set of features which progressively grows over the course of the test. Each set of features included in the test corresponds to a different view. This allows us to ascertain whether there is an improvement in the classification by including additional information acquired from the other views. We used standard precision-recall curves to check the effect of adding features.

4.4. Publications

We present our method as part of three publications. The first work includes a concept test using our approach to solve a simple and known detection problem. We evaluated the projection method and an ensemble approach to classify flaws in aluminum die casting. The second article

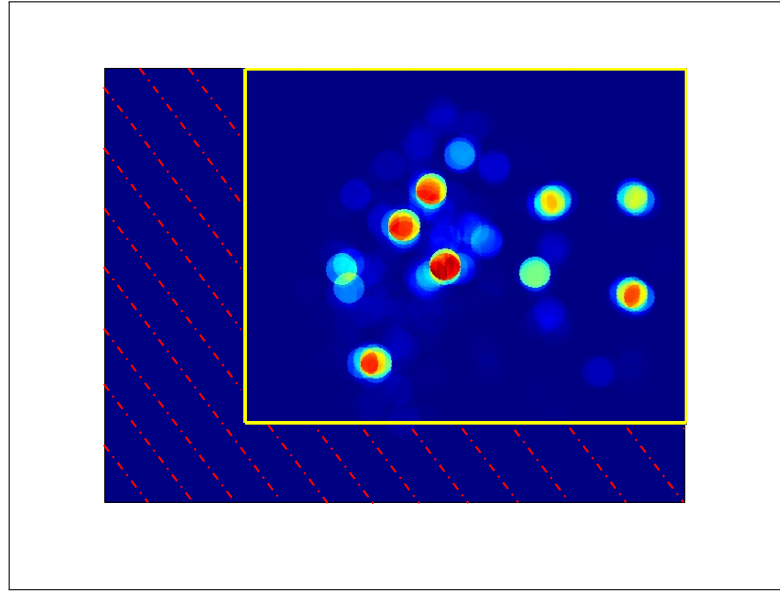


Figure 4.6. Example of heat map used to analyze the most frequently visited areas. This map shows a top-view of the classroom used for our experiments. Pseudo color indicates the number of detections at each location. Red areas are equivalent to a high number of detections, blue areas point a low number of detections. Similar as we describe in Fig. 3.5, the space of interest S is limited by the yellow square. The dashed region marks the space that we do not take into account during detection.

describes an application of the ensemble of features and ensemble of classifier methods to head representation and classification. And the third publication describes our head detection method, including most of results presented in this thesis.

4.4.1. Flaws Detection

Due to security requirements, the car industry requires that 100 percent of the aluminum parts included in vehicles be inspected. X-rays are the main non-destructive testing method used to identify flaws within an object that are undetectable to the naked eye. Although human inspectors are hard to replace, they can only meet these requirements for short periods of time. There are human factors, such as monotony, tiredness, eye stress, and loss of concentration that make this task sensitive to errors (Boerner & Strecker, 1988). Automated inspection using radiographic images has been made possible by incorporating image processing techniques into the process. There are successful mono-focal approaches (Boerner & Strecker, 1988; Anand, Kumar, et al., 2006) which demonstrate the ability of those autonomous system to assist humans in this task. However, false positives have to be kept to a minimum due to cost constraints. Thus, multiple view approaches

are an important tool for detecting and reducing false alarms. Multiple view approaches work in calibrated (Mery et al., 2005, 2002) and uncalibrated scenarios as well (Carrasco & Mery, 2006). Both multiple view approaches consist of flaws in the segmentation of each image, searching for matches in the images, and tracking the flaws in the images. In general, those approaches require effective segmentation and matching algorithms. We propose an automated inspection using our sliding-box approach. The system has three main parts: multiple view projection, feature extraction, and classifier for sequences. The classifier learns the flaw structures from simulated flaws. Thus the system uses all of the evidence that is made available by the multiple views. We combine the information gathered without having to search for matches. The classifier then determines whether an image sequence belongs to flaws/no-flaws classes in the detection process, as shown Fig.4.7. In this study, we use 72 views of the same aluminum wheel. We arrive at two important conclusions: 1) simulated flaws can be used to train classifiers used in these applications due to the fact that real flaws are rare events in industrial manufacturing processes and 2) simultaneous combination of information from different viewpoints using sliding-boxes is a robust approach to flaw identification. The results and experiments were described in (Pieringer & Mery, 2010), and will be included in the Appendix A.

4.4.2. Head Modeling

Object detection has attracted great interest of researchers in the computer vision community. Although machine learning approaches has been successful in this task, there are still significant challenges to solve in order to achieve data association, and including information from various points of views. We propose a multiple-view classification approach to bring a gap between advances in machine learning based object detection and multiple view geometry. The key idea is to classify an image sequence of corresponding parts of an object. This scheme allows us to solve problems related to correspondence throughout cameras, and to enhance the detection models with compounded features. This article describes our approach applied in human head modeling by integration of visual information. The experiments demonstrate that our technique improves 2D state-of-art classifiers, using same training conditions. These results are promising and show that our approach can be use effectively to detect objects using multiple views. The results and experiments were described in (Pieringer et al., 2012), and will be included in the Appendix B.

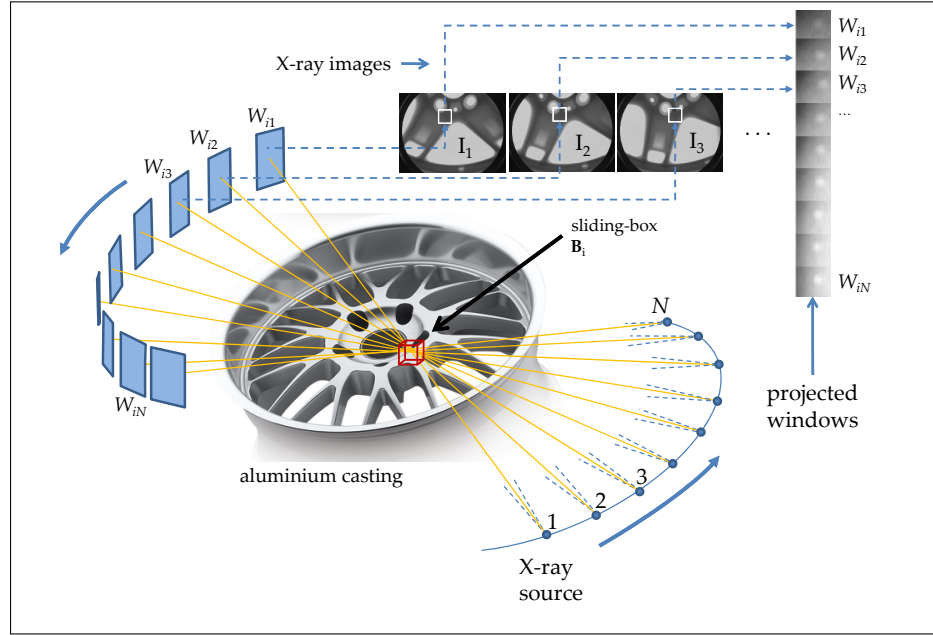


Figure 4.7. Concept testing for our approach, applied to flaws detection in aluminum die casting. We build a flaw sequence using 72 images in a calibrated multiple view system. The classifier learns from simulated flaw sequences and identifies flaws on real images.

4.4.3. Head Detection Using Sliding-Boxes In Multiple Views

Sliding-Window detectors have attracted great interest among researchers in the computer vision community because of the advantage they offer of avoiding segmentation problems during detection. significant progress in detection is due in part to the use of powerful machine learning models, new and more informative visual features, and part-based models which cope with object variability (Viola et al., 2005; Dalal & Triggs, 2005; Felzenszwalb et al., 2009; Yang et al., 2009; Girshick et al., 2014; Dean et al., 2013). Although there have been some improvements, overall performance is still poor (Dollar et al., 2011). A common denominator of these techniques is that they rely mainly on statistical learning methods that exploit image-intensity information to capture object appearance features. Their goal is to reveal visual spaces where visual similarities carry enough information to achieve robust visual recognition. An important limitation of appearance-based approaches is that they do not incorporate relevant geometric information that can provide important, useful spatial cues such as the real size of the object to be detected, depth, and likely spatial appearance location. Some notable exceptions with promising results are Salas and Tomasi

(2011); Spinello and Arras (2011); Espinace et al. (2013); however, these approaches require additional hardware to recover the spatial information. Detection using one camera is suitable when there is mild occlusion, but if there is heavy occlusion multiple views help to improve final detection (Mittal & Davis, 2003; Khan & Shah, 2006; Eshel & Moses, 2010; Liem & Gavrila, 2013). Geometric techniques exist for establishing relationships across views in a camera system, providing useful ways of combining and integrating this information (Hartley & Zisserman, 2003; Szeliski, 2010). However, there are practical constraints when the appearance of the target object changes drastically from one viewpoint to another. In general, we observe that single view approaches to object detection mainly *i)* use a sliding-window at various scales to compensate scale changes of the object target class in images, resulting in false positives due to hallucinations at several scales; and *ii)* do not take into account useful 3D information such as real sizes of people or objects, and the positions in which they are likely to be found in the scene.

We propose a generalization of the sliding-windows approach to 3D cases, which we call sliding-box. This approach works on calibrated multiple view configurations where we have various viewpoints of the same scene. This calibrated system allows us to run a sliding-box in world coordinates and project this box in its corresponding position on each image. We can also search in the correct scale and location, improving the processing limitations of the sliding-box. Furthermore, it allows us to create appearance models from various viewpoints to improve detection.

We apply our approach to head detection as the head is useful when the rest of the body is occluded. Experiments show that our framework improves detection performance by 10% of average precision-recall as compared to the optimal view of state-of-the-art 2D methods in our datasets. These results suggest that our approach can be used effectively to detect objects in multiple view systems, improving detection performances achieved by 2D detectors in isolation. We include a preliminary version of this article in Appendix C.

Chapter 5. EXPERIMENTS AND RESULTS

In this chapter, we describe the experimental results of applying the sliding-box approach in people detection. We present the qualitative and quantitative results of the algorithm, and compare our results to those of a state-of-art 2D multi-scale detector. We evaluate algorithms according to the methodology presented in Chapter 4, in which the qualitative results showcase the practical advantages of our approach, and precision-recall curves are used to measure and compare the performances of detection algorithms based on the qualitative results. We run detection tests using two real video sequences acquired within a classroom environment in order to validate our hypothesis and test the performance of our approach. Both test datasets were manually labeled to identify the ground truth elements. In general, our experiments confirm that using multiple views complements the data in the vision system and increases the overall detection performance. Detailed information about implementation and results will be discussed in the next four sections.

5.1. Implementation Details

During testing, we run the sliding-box in a fully labeled test dataset T . The sliding-box movements are geometrically limited to region of interest S , in which we locate true head hypotheses and discard those that fall out S . In our implementation, we use a dense grid mesh with steps of 5cm to move the box. This exhaustive analysis produces $\sim 60,000$ boxes, depending also on the size of the region of interest. We drastically reduce the number of hypotheses by applying the 3D focus-of-attention procedure. During our experiments, we consider $N_{matches} = 2$ correspondences to compute the focus-of-attention. Our sliding-box runs along the regions identified by the focus-of-attention. The box \mathbf{B}_i is projected onto the images as W_{ij} , and then cropped and rescaled to 64×64 pixel, in order to create a collection of aspects as is shown in Fig. 3.4. We describe each element W_{ij} using the uniform LBP features over a 3-level pyramidal decomposition, obtaining a total of 21 feature blocks per object instance. This is the number of levels recommended to avoid overfitting (Bosch et al., 2007). Although HOG and LBP descriptors are state-of-the-art low level features, LBP performs better than HOG in early experiments. We explain that difference because head aspects present different textures in different poses. Some articles describe this phenomenon in detection, explaining that LBP cope better with color or textured information (Mu et al., 2008; Wang et al., 2009). We use the VLFeat library to compute LBP features (Vedaldi & Fulkerson,

2010). In the course of detection, the multiple view model evaluates each hypothesis considering all the possible alignments in order to select the most confident set of aspects, as we described in Section 3.4.4. In both cases, we use support vector machines (SVM) with linear kernels as classifiers (Cortes & Vapnik, 1995). We compared linear and RBF kernels, and the former provided more reliable results for different choices of parameters. As stated in Yang et al. (2009), linear kernels perform better in visual applications with large and sparse descriptors. The next two sub-sections describe how we trained both classification approaches for projection sequences. As we mentioned previously, both are independent and exclusive of each other. Our implementation takes an average of six minutes to classify 1,500 boxes using MATLAB over Ubuntu-Linux and an AMD Phenom II X4-925, 4GB RAM, 2.8GHz computer.

5.2. Detection Performance

The main goal of our method is head detection in order to determine people’s location. We study the detection performance of our sliding-box approach using that of a state-of-art 2D multi-scale detector as baseline. We compare the performance of our two approaches using as a baseline the Deformable Part-based Model (DPM) (Felzenszwalb et al., 2009). In order to obtain a fair comparison, we use epipolar geometry to filter out false positive detections of the DPM detector, as we explain below. We performed experiments using test datasets described in Section 4.2. Performance was evaluated following the structure set out in the PASCAL challenge using precision-recall (PR) curves (Everingham et al., 2010).

We train the DPM algorithm using our head dataset. After training, this model allows us to detect heads in the testing dataset, as shown Fig. 5.1. First, we run the 2D detector for each camera independently. Then, detections in one view are tracked through the camera system using epipolar lines in order to establish corresponding detections from all views.. We considered a detection in camera i to be true if a detection in camera j appeared near to the epipolar line computed using the detection in camera i , such that $\mathbf{l}_j = \mathbf{F}_{ij} \cdot \mathbf{m}_i$, where \mathbf{F}_{ij} is the fundamental matrix between views i and j , and \mathbf{m}_i is centroid of the detection in camera i . Otherwise, the filter discards detections that could not be tracked through N_{match} views in the camera system. To make the evaluation between multiple view methods and multi-scale methods fairer, we also include prior knowledge during 2D detection, such as the maximum and minimum bounding-box size, and likely location for detection

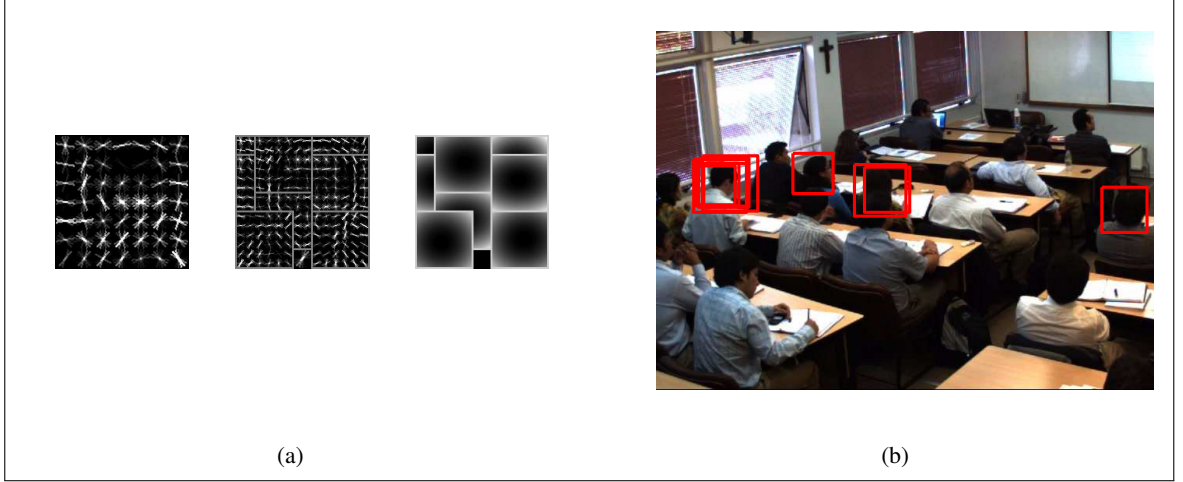


Figure 5.1. Summary of the head model that the DPM learns after training. (a) Example of the model, its gradient map, and parts. (b) Example of detections after applying the model over test images.

in the image. We evaluated these approaches using the methodology proposed in VOC Challenge PASCAL (Everingham et al., 2010), where detections within an overlap area of the ground-truth counts as correct detections. Detections outside of this area counts as false detections.

In order to evaluate our approach we calculate the performance of our method using a modified version of the PASCAL method. Since we know spatial information such as box size, we simplify the function evaluation calculating the Euclidian distance between the detected sliding-box SB_{dt} and the ground truth box SB_{gt} . Both, SB_{gt} and SB_{dt} , form a potential match if the distance is lower than the overlap radio $r_{overlap}$,

$$distance \doteq \|SB_{dt}, SB_{gt}\| < r_{overlap}. \quad (5.1)$$

The overlap radius depends on prior spatial information. In our experiments, we consider $r_{overlap} = 180[mm]$. This value is the average radius which produces an average overlap of 50% for 2D detection according to the PASCAL criterion. We can also explain this value geometrically. With the exception of specific cases in which two or more heads are very close to one another, heads are generally located at the center of the body and are therefore geometrically separated from other heads at a minimum distance equivalent to the space occupied by the whole body. Based on this assumption, if we consider a correct head detection that match with its ground-truth any other

overlapped detections surrounding it within a radius lower than $r_{overlap}$ there will be the same. We apply the DPM detector and DPM detector variation using epipolar geometry for the four views of the dataset. After detection using these 2D detectors, each view has its own performance curve. Comparison includes precision-recall curves of our approach compared to the performance curve of these 2D detectors corresponding to the view with the highest AP value.

We evaluate our detection approaches in both datasets using suppression the methods described in Chapter 4: Weighted Mean Shift suppression (WMS-NMS) and Local Maximum Searching suppression (LMS-NMS). For the purpose of the evaluation, we considered detections that match with ground-truth heads present in four views and ignored from the evaluation detections that match with the ignored ground-truth heads. Figures 5.2 and 5.3 show that the one-vs-one Ensemble of Classifier (OVO-EC) model using WMS-NMS outperforms the Ensemble of Features (EF) and the one-vs-all Ensemble of Classifier (OVA-EC). These results show that suppression affects the overall detection performance. These variations are more noticeable in dataset *sq-01* where pose changes may to influence the detector performance. WMS-NMS represents the overlapping boxes using a weighted mode that finally fits better with the real ground-truth location. Our approaches generally reach better performances than the DPM detector, producing similar recall rates but allowed for higher rates of precision at the tail of the curves, as compared to the 2D detections. This proves that using combinations of viewpoints may improve detection performance. We conclude that mean-shift suppression generally contributes to improving the location of final detections even using the different classifiers. On the other hand, LMS-NMS based uses a sampling strategy that not always fit with the real location of heads according to ground-truth.

Specifically in dataset *sq-01*, the detector based on OVO-EC and mean-shift suppression perform better than other approaches, including 2D detectors. Although detectors using LMS-NMS achieve good levels of performance, DPM detector perform the best, as shown in Fig. 5.2a. As we have mentioned, OVO-EC performs the best and showed improved precision on the last part of the curve, though this is less noticeable. Though both versions of DPM detector performed at similar levels, we note that DPM detector using epipolar geometry presents a small improvements in precision over recall of 0.85 where it could to eliminate some false detections, as shown in Fig. 5.2b and Fig. 5.2a. Nonetheless, this improvement is not enough to produce an overall enhancement to outperform the DPM without epipolar geometry. The results show that epipolar geometry contributes

to eliminating false positives, as we stated, but in this case it does not allow us to recover missed detections in any viewpoints.

The results of dataset *sq-02* show behaviors similar to dataset *sq-01*. The overall detection performance presents few variations among the three multiple view approaches. They are all quite similar in terms of precision-recall. They yield an AP value of $[0.73 - 0.74]$ regardless of the suppression procedure, as show in 5.3a and Fig. 5.3b. The camera viewpoint may improve the location estimation during focus-of-attention computation and eliminate the influence of the NMS. We observe classifiers performed on similar levels due to head pose because people pose naturally facing the two frontal cameras. This alignment is similar to the samples used to train the model, which means that all classifiers behave in a similar fashion. We also note the same improvements in precision after recall of 0.5, which means that these approaches outperform 2D detectors. At this recall point, the performance of both 2D detectors decreases in quality. We do not observe detection improvements in the DPM detector based on epipolar geometry.

Tables 5.1 and 5.2 show the confidence interval for detection in order to demonstrate the repeatability and accuracy of our detectors. We use a procedure that is quite similar to the one that we use to generate precision-recall curves. The procedure first separates random sets of test instances into k -folds, and computes the AP value for each fold. Finally, it uses theses k -AP values to compute the average and the confidence intervals using a t-Student Test at 95% of confidence level. We set $k = 10$ that is the standard value used in cross-validation evaluations. The results show little variation of AP values in LMS-NMS and WMS-NMS suppression methods. This test find averages AP values that were close to the AP values presented previously in tests for overall detection performances. Most intervals present a significance of over 95% of confidence level and error below 0.05. The detection results show that WMS-NMS improves the final detection in terms of quality. As we stated previously, the WMS-NMS performs better than the LMS-NMS procedure in terms of how much it helps improve the location and rejection of overlapped sliding-boxes. However, there is a decrease in precision due to the low detection thresholds because mean-shift computes the local mode using spurious hypotheses, as show in Fig. 5.3b.

Figures 5.4a and 5.5a show a heat-maps plots that summarize locations of heads within S in sequences during detection. Heat-maps are the cumulative log-normalized sum of frames in the sequences (Dollar et al., 2011). We use these plots to visualize the effects of each classification

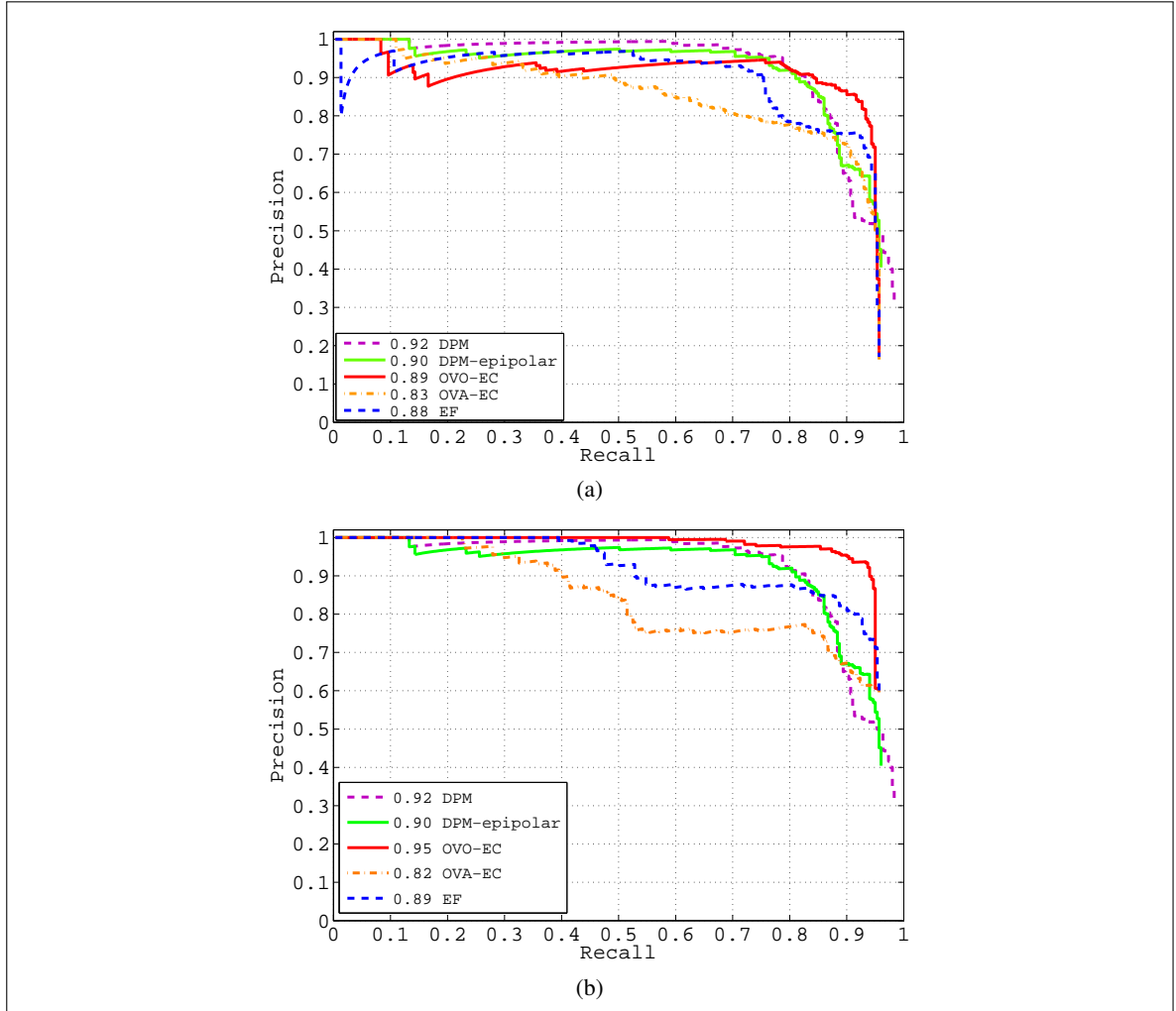


Figure 5.2. Precision-Recall curves comparison of detection performance in sequence *sq-01*. We compare performances of our methods EF, OVA-EC, and OVO-EC versus performance of single view DPM and the epipolar version of DPM. In general we report high performance level of our OVO-EC approach. The overall performance is represented using the AP value. (a) Performance evaluation using LMS-NMS. Although detectors performed well, the DPM performed best overall. Our approach based on a one-vs-one ensemble of classifier presented the best performance. We note an improvement in precision on the last part of the curve that allowed to our method to maintain its level of precision at the same recall rates as DPM. (b) Performance evaluation using NMS based in mean-shift procedure. In this case, our approach based on a OVO-EC of classifier performed best. We observed the same improvement in precision on the last part of the curve, but in this case it was more noticeable. In both curves, DPM using epipolar geometry showed marginal improvement after recall of 0.85. This may be due to the way that epipolar geometry helps to eliminate false detections.

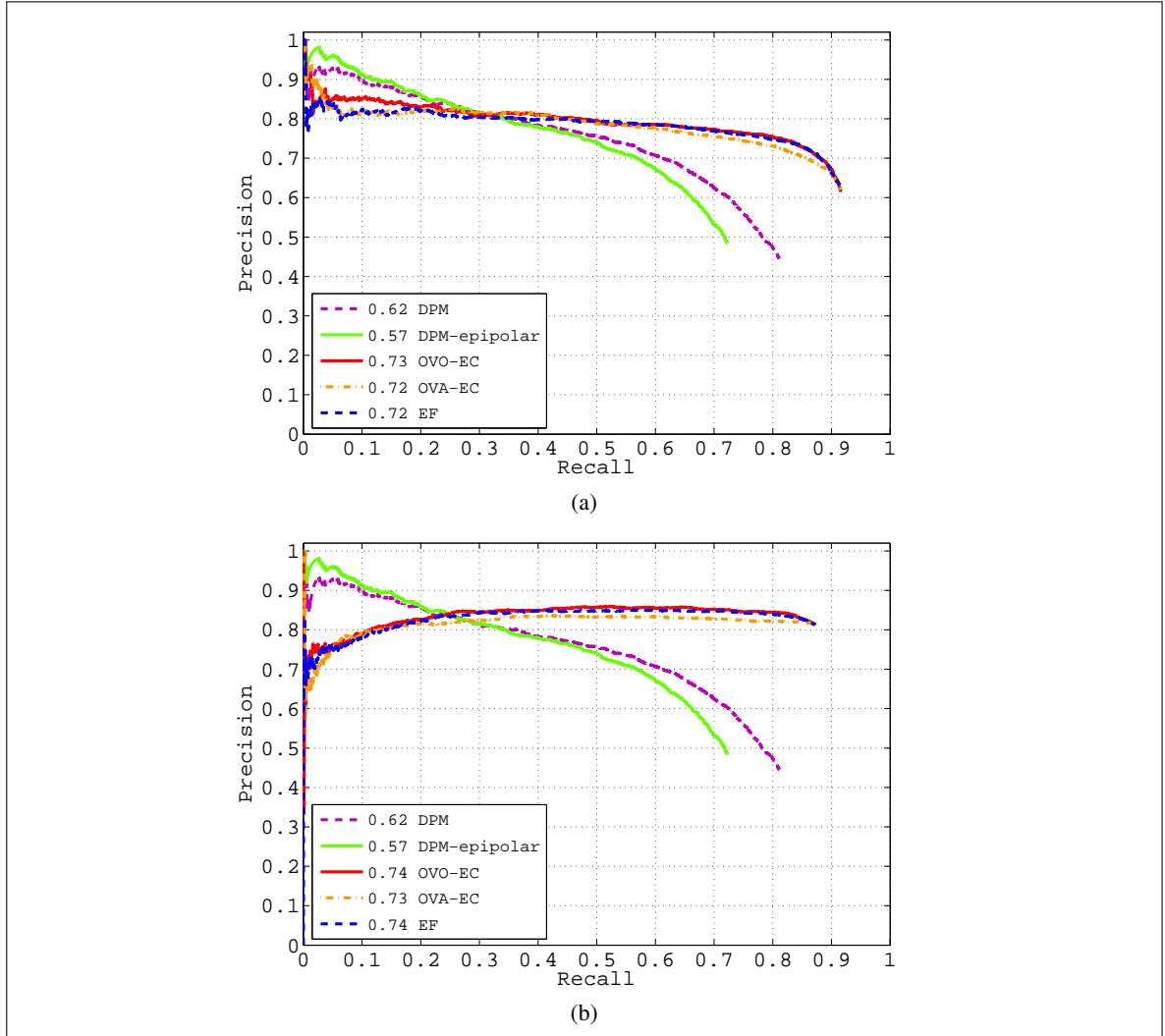


Figure 5.3. Precision-Recall curves comparison of detection performance in dataset *sq-02*. Comparisons include our three methods EF, OVA-EC, and OVO-EC versus performance of single view DPM detector and the epipolar version of DPM. In general we report best performance level in the OVO-EC approach. The overall performance is represented using the AP value. (a) Performance evaluation using NMS based on greedy procedure. The three multiple view detectors performed at similar similar levels. All of them outperform the DPM detector in its two versions. Our approach based on a OVO-EC yielded the best performance as in *sq-01*. We note improvements in precision on the last part of the curve which allowed our method to maintain a level of precision at the same recall rates as DPM. (b) Performance evaluation using NMS based on mean-shift procedure. Our approach based on a one-vs-one ensemble of classifiers performed best. The same improvements in precision are present. In both curves, DPM detector using epipolar geometry does not show an improvement. This may be due to the fact detections near to same epipolar lines are eliminated.

Table 5.1. Repeatability analysis of our three approaches in dataset *sq-01*. Results show little variation after detection. All methods show significance over the 95% of confidence level.

	LMS-NMS			WMS-NMS		
Method	Avg. AP	σ	Confidence Interval	Avg. AP	σ	Confidence Interval
OVO-CE	0.905	0.061	[0.861 - 0.948]	0.953	0.025	[0.935 - 0.971]
OVA-CE	0.836	0.058	[0.794 - 0.877]	0.855	0.058	[0.813 - 0.896]
FE	0.884	0.042	[0.854 - 0.913]	0.926	0.037	[0.899 - 0.952]

Table 5.2. Repeatability analysis of our three approaches in dataset *sq-02*. The results show little variation in detection that passes the 95% of confidence level.

	LMS-NMS			WMS-NMS		
Method	Avg. AP	σ	Confidence Interval	Avg. AP	σ	Confidence Interval
OVO-CE	0.747	0.053	[0.709 - 0.785]	0.754	0.034	[0.730 - 0.778]
OVA-CE	0.733	0.062	[0.688 - 0.788]	0.738	0.043	[0.707 - 0.769]
FE	0.739	0.053	[0.701 - 0.777]	0.751	0.052	[0.714 - 0.788]

approach and suppression method. In general, the results show that WMS-NMS provides more accurate outputs than the smoother outputs of the LMS-MNS. This suppression affects the final performance. The geometric information about the scene also allows us to limit the analysis to the space defined by the yellow square S , and to discard the crosshatch area because it does not contain candidate points. Detection results show that both ensemble approaches look similar to each other, but we note that ensemble of classifiers using one-vs-one strategy produces results that are a better match with the ground-truth, particularly using WMS-NMS procedure, as shown in Fig. 5.4f and Fig. 5.5f. Although the ensemble of features and the ensemble of classifier using a one-vs-all strategy produce similar results to the ground-truth, we detected the presence of artifacts that can reduce the quality of final detections, as shown Fig. in 5.4 and Fig. 5.5.

Visually, we also can compare the best view of PDM detector based on the AP achieved in the PR curve to the detections yield by our approach in the same view. We set the detection threshold to 0 for practical reasons. Figures 5.6 and 5.7 show detection provides by all methods. We note that our approach retrieves detections missed by DPM. We display detections in red boxes and ground-truth elements in green boxes. Although our approaches add missed detections in one view, they also hallucinate some noisy detections due to an incorrect combination of data. This example also presents noisy detections after using LMS-NMS, especially for OVA-EC and FE approaches. The

OVO-EC presents the best quality of detections in accordance with centering of the bounding-box, as we stated in experiments of precision-recall curves.

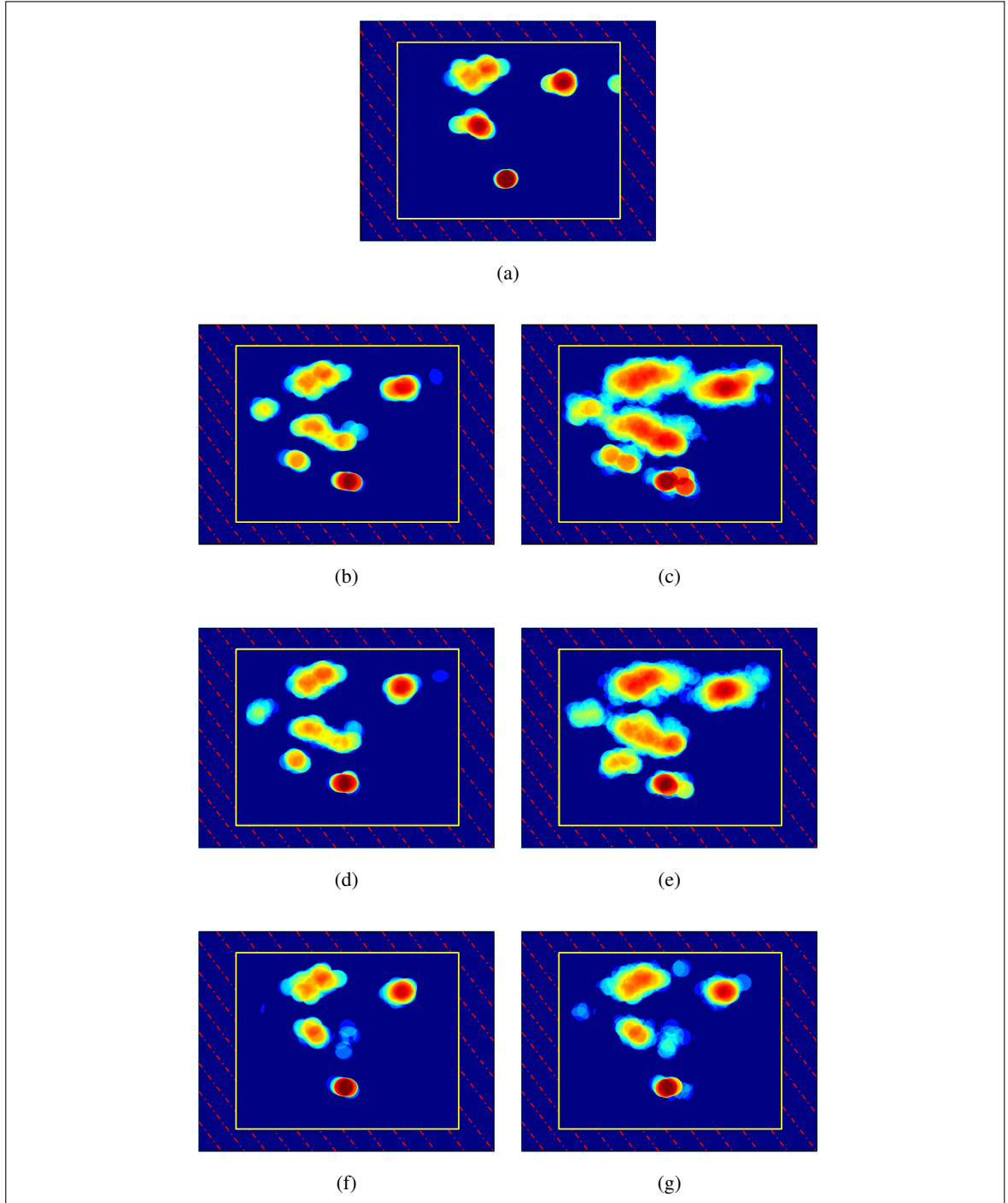


Figure 5.4. Heat-map of detections using our three detection approaches in sequence *sq-01* after applying WMS-NMS and LMS-NMS. Left column are detection using WMS-LMS and right column are detection using LMS-NMS. (a) Heat-map of the ground-truth. (b) and (c) Heat-maps of detections using the EF approach. (d) and (e) Heat-maps of detection using OVA-EC. (f) and (g) Heat-maps of detection using OVO-EC.

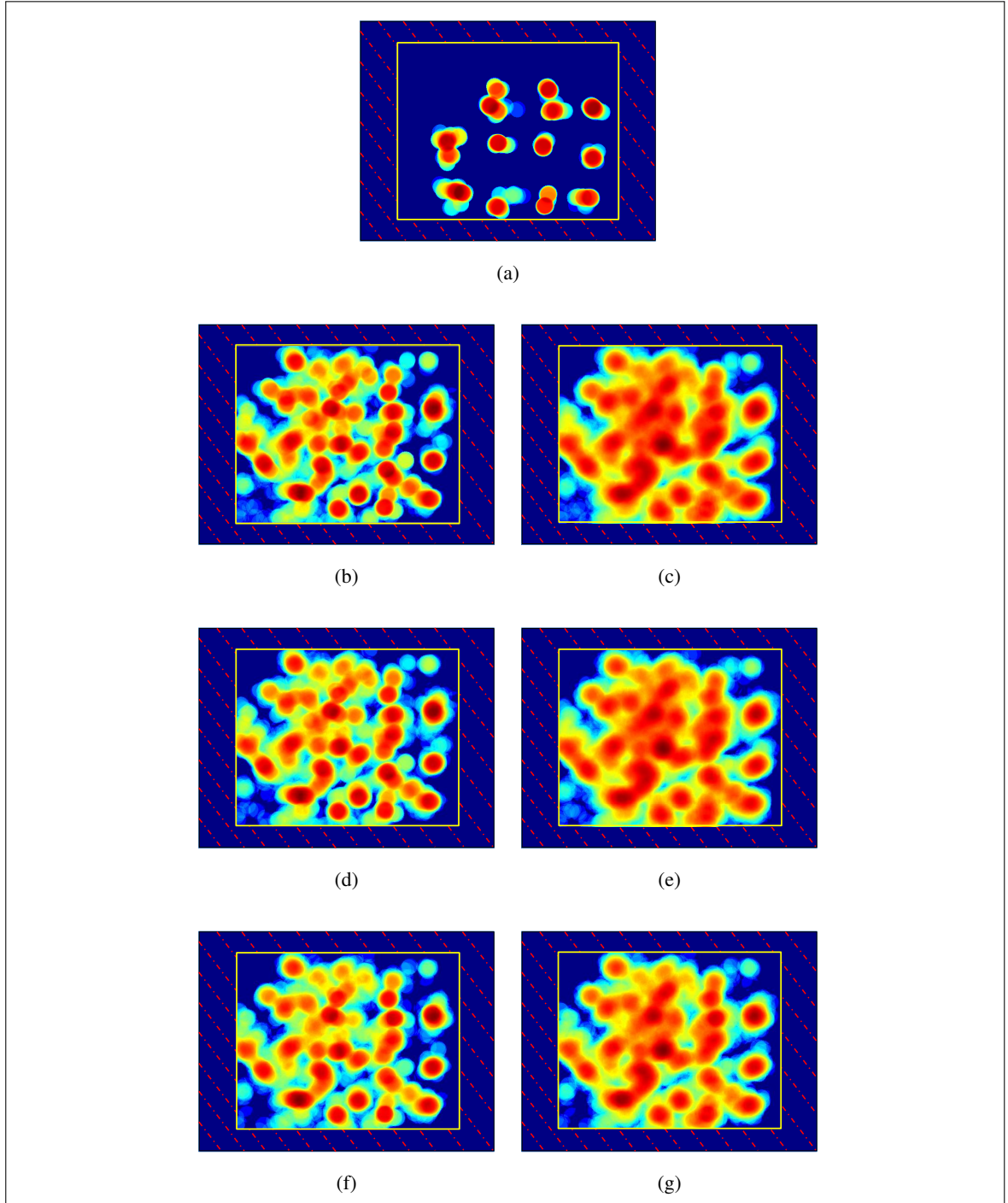
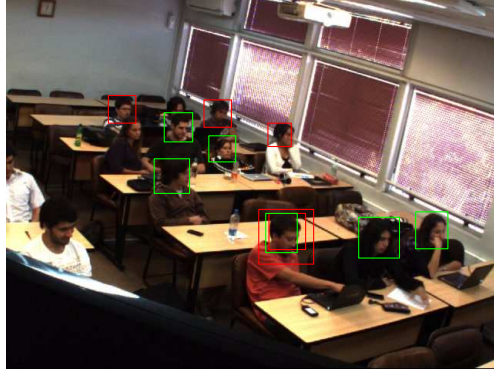
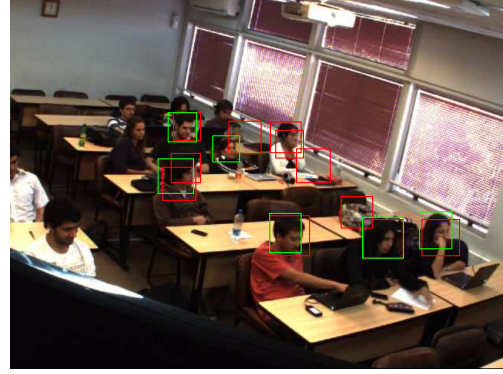


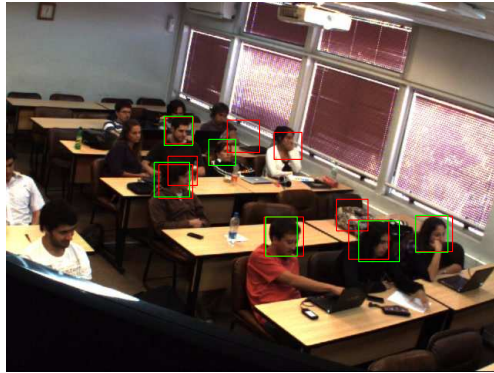
Figure 5.5. Heat-map of detections using our three detection approaches in sequence *sq-02* after applying WMS-NMS and LMS-NMS. Left column are detection using WMS-LMS and right column are detection using LMS-NMS. (a) Heat-map of the ground-truth. (b) and (c) Heat-maps of detections using the EF approach. (d) and (e) Heat-maps of detection using OVA-EC. (f) and (g) Heat-maps of detection using OVO-EC.



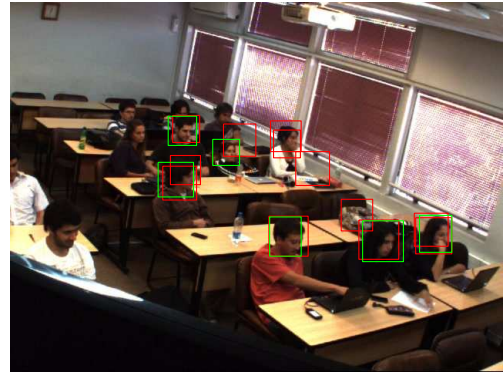
(a)



(b)

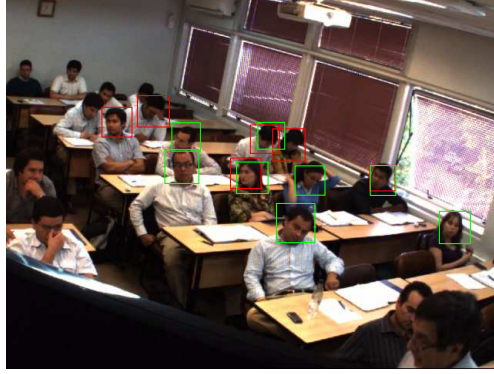


(c)



(d)

Figure 5.6. Detections comparison using WMS-NMS. (a) Detections provided by DPM detector. (b) Detections generated by OVA-EC approach. (c) Detections yielded by OVO-EC approach. (d) Detections generated by applying the EF method. We display detections in red boxes and ground-truth heads in green boxes. The results show that our methods can retrieving heads that the 2D detector missed. Although we retrieve new heads, we also add some noisy detections.



(a)



(b)



(c)



(d)

Figure 5.7. Detections examples using NMS based on mean-shift. (a) Detections provided by Latent-SVM detector. (b) Detections generated using OVA-CE. (c) Detections yielded by OVO-CE. (d) Detections yield by applying FE method. Detections appear as red boxes and ground-truth heads as green boxes. The results show most of the heads missed by the 2D detector were retrieved by our multiple view detector. Although we retrieve new heads we also add some noisy detections due to hallucinations of the classifier.

5.3. Enriched features

We analyze the influence to train the model β of our best approach varying the number of cameras. This test allows us to evaluate the impact of using different visual sources and the power of combining features from various viewpoints of the same object. First, we train the model β only using one camera. Second, we progressively re-train the model adding one camera at a time. We perform this analysis in terms of per-window performance using AP value without considering the suppression procedure after classification in order to evaluate the discriminative power of our multiple view classifier (Dalal & Triggs, 2005). Figure 5.8 indicates that the three methods show an increase in performance as we add a new camera. When we train using a single camera, we report a low rate of average precision-recall (AP) that proves the classifier using single viewpoint generates weak models. This demonstrates that complementary information enhances the model β and improves the classification. We also found that EF method shows a noticeable performance gap between two and three cameras, as shown in Fig. 5.8a. The same performance improvement is present in both ensemble of classifiers approaches, but the transition between the step is smoother than the ensemble of features, as shown Fig. 5.8b and Fig. 5.8c. The three classification schemes show a marginal improvement when the classifier learns from three and four cameras.

5.4. Focus-of-Attention

The focus-of-attention mechanism and the spatial context allow for object target class be correctly located in applications such as train stations, restricted areas, and points of access. In Chapter 3, we explained that sliding-box runs across the 3D space in order to predict the location of heads. During this process, the algorithm must process a massive number of windows. We compute the spatial focus-of-attention in order to decrease this amount of information to be processed. This procedure generates spatial hypotheses of the real location of heads. In our implementation, we use the output of an interest-point detector to estimate the spatial position of latent candidates by triangulation (Hartley & Zisserman, 2003; Szeliski, 2010). In this way, our attention mechanism produces spatial hypotheses in the most likely locations for finding heads.

In our implementation, we use the detections provided by the 2D detector that works as a specialized interest-point detector. The focus-of-attention procedure does not have restrictions in terms of this type of interest-point detector. However, we prefer to use these detections because of the

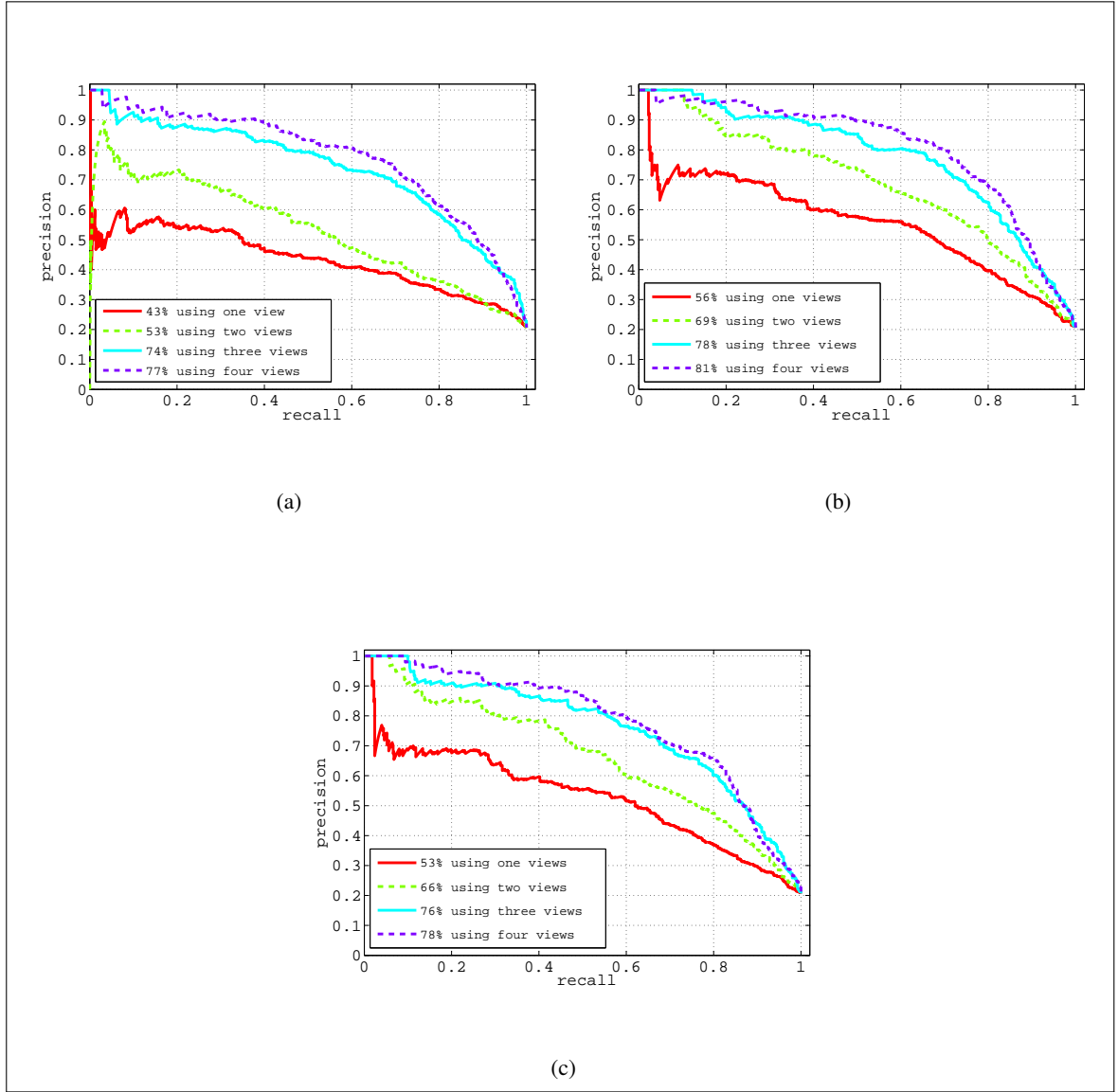


Figure 5.8. Performance comparison in per-windows classification. The curves show performance evolution as we add information from the four visual sources of our camera system. (a) precision-recall curve of ensemble of features. This method presents an improvement in performances after the third camera is included, which is the view that includes information about the rear of the head. We conclude that the additional viewpoint contributes to improving the performance. (b) The precision-recall curve of ensemble of classifiers using a one-vs-all strategy. (c) The precision-recall curve of ensemble of classifiers using a one-vs-one strategy. These second method presents improved performance as we add a new viewpoint. It is smoother that the improvement presented by ensemble of features. Overall performance is represented using the AP value. High AP values are yield by all methods when we use the four cameras.

Table 5.3. Summary of the burden reduction due to the spatial focus-of-attention. Although there is an overhead due to applying an interest-point detector, it is always better use this procedure than perform an exhaustive search across the region of interest.

Sequence	Total no. of Boxes	Percentange of Burden Reduction Varing the Detector Threshold						
		-1.0	-0.7	-0.4	0	0.4	0.7	1.0
<i>sq-01</i>	92,160	85 \pm 3.5	95 \pm 1.5	98 \pm 0.65	99 \pm 0.5	99.5 \pm 0.3	99.7 \pm 0.27	100 \pm 0.15
<i>sq-02</i>	65,536	92.5 \pm 3	96.3 \pm 1.6	98.3 \pm 0.4	99.3 \pm 0.5	99.4 \pm 0.5	99.6 \pm 0.3	99 \pm 0.2

implementation goes beyond the scope of this thesis. Figure 5.9 shows that the total number of hypotheses and therefore the maximum recall that we can achieve depends on detector confidence, *i.e.*, the higher the confidence threshold, the smaller the number of spatial hypotheses and the chances of recovering the total number of heads. We use a sensitivity analysis to show the maximum recall achieved for each video sequences. The maximum recall in sequence *sq-01* is between -0.4 and 0 of detection confidence. In sequence *sq-02* the maximum recall is in the range of -0.7 and -0.4 of the confidence. Table 5.3 summarizes the burden reduction by using focus-of-attention. The average burden reduction is 95% at a threshold that allows us to recover all detections in the region of interest S . Although there is an overhead due to applying an interest-point detector, and a trade-off between the number of spatial hypotheses and the detector confidence, it is always better to use focus-of-attention than perform exhaustive searching across the detection region. Figure 5.10 shows examples of focus-of-attentions provided by our algorithm. The examples in the left column show results using a high confidence value of the 2D detector. This examples show the decrease of head hypotheses, and therefore the recall of the detection algorithm. Examples in the right column show results using a low confidence value. We note that most of the ground-truth elements match with the spots defined by our attention mechanism.

5.5. Using of Geometrical Cues to Detection

We stated in Chapter 1 that our approach offers advantages and contributions related to the knowledge about prior information about the target object classes. We perform a qualitative evaluation comparing the windows used by two detection method: sliding-window detectors, and our multiple view detector. We pick two sets of a random sample of 60 windows in both detection approaches. Windows used with the 2D detector covers various scales trying to predict the size of the object, which produces hallucinations at scales with higher resolutions, Fig. 5.11a. Our approach

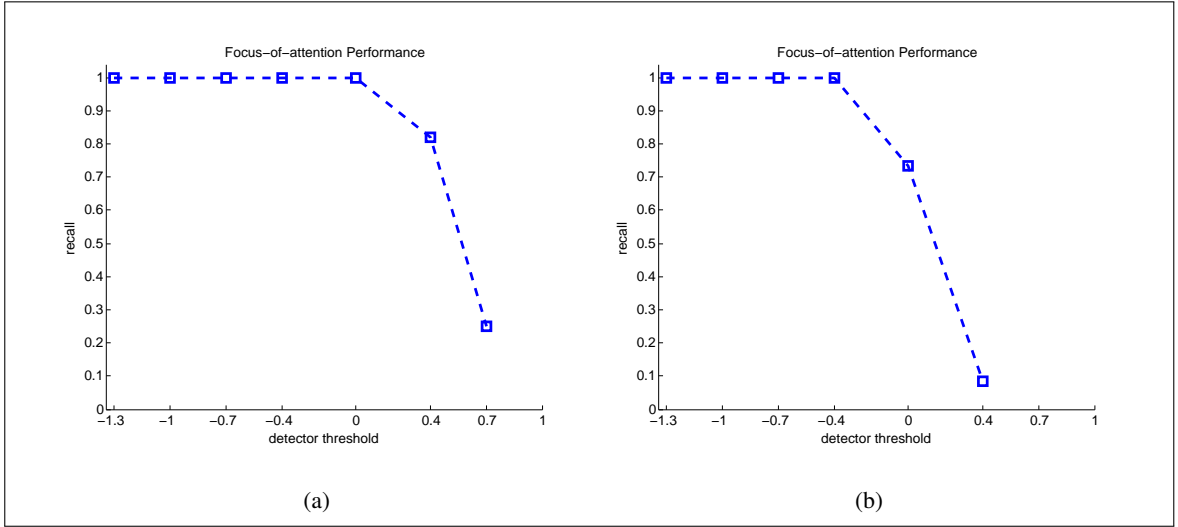


Figure 5.9. Sensitivity analysis of confidence vs recall due to use of the spatial focus-of-attention procedure. There is a trade-off between detector sensitivity and the maximum recall reached upstream of the classifier. If we set a low sensitivity for the detector, we increase the ability of the focus-of-attention to retrieve all detections in the scene. Even though a low confidence threshold produces more spatial hypotheses and increases the burden for other procedures, it is always better use the focus-of-attention than to run the sliding-box across the entire region of interest \mathcal{S} . (a) shows sensitivity for sequence *sq-01* where the maximum recall appears to fall between 0 and -0.4 of detector confidence. (b) shows sensitivity for sequence *sq-02* where the maximum recall appears to be between -0.4 and -0.7 of detector confidence.

can control the size of the windows and the areas analyzed. This allows us to generate more informative windows for the classifier and limit the analysis to more likely location, as shown in Fig. 5.11b.

The main task of the multiple view classifier is to discard boxes that include projections which belong to the background. However, we note that the models intrinsically also perform the task of aligning the collection of aspects that they learnt during training. In Fig.5.12, we show three sets of four candidate boxes. The first and second sets, shown in Fig.5.12a and Fig.5.12b, belong to the head class; the third set, in Fig.5.12c, belongs to the background class. Once the algorithm evaluated each hypothesis, higher scores were always given to the best alignments, which had the best scores and were therefore selected by the algorithm. All of the scores in the third set c) are strongly negative, and therefore all were assigned to the background class.

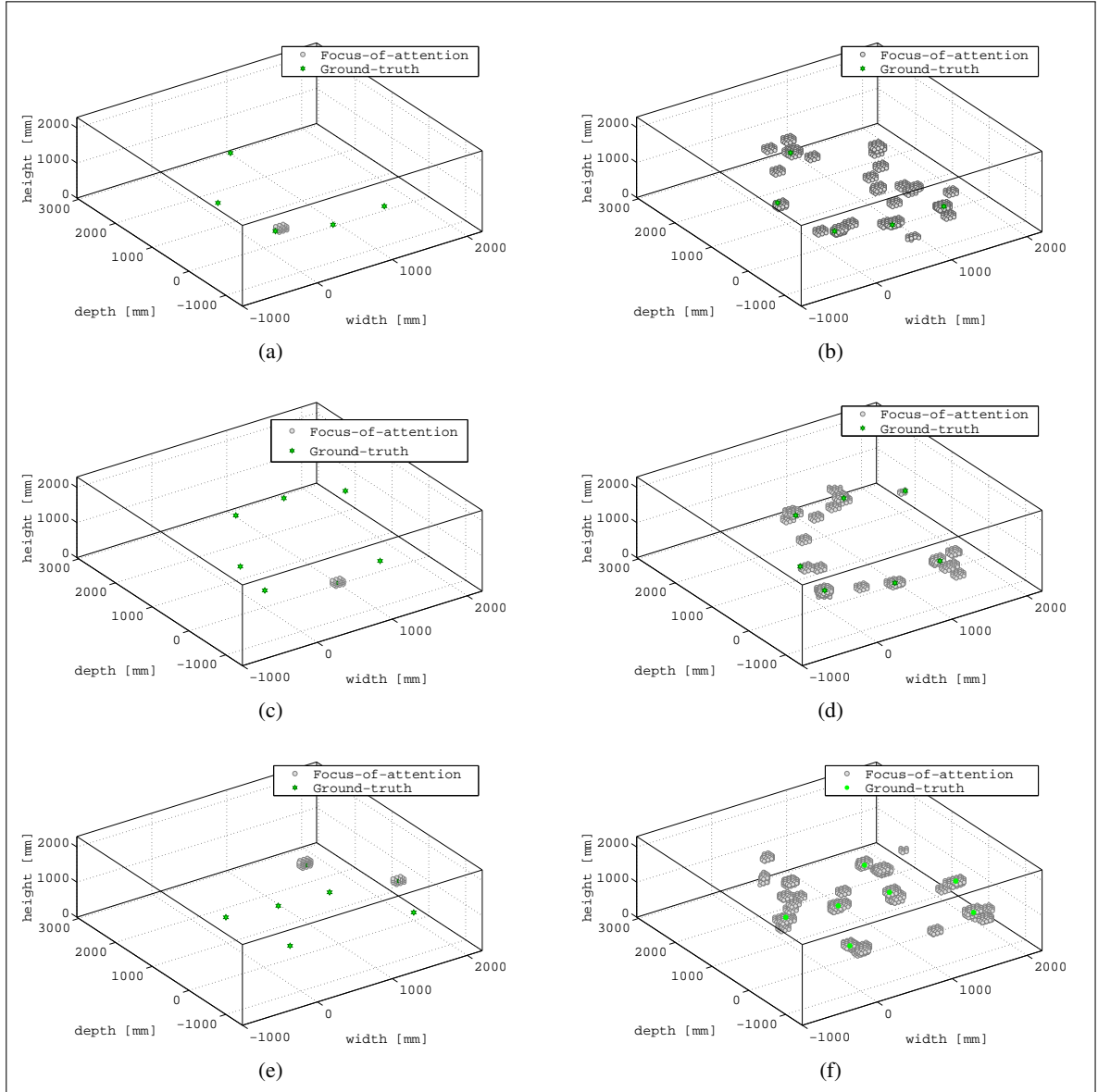


Figure 5.10. Examples of focus-of-attention computed within the region of interest S using the outputs provided by an interest-point detector. The head hypotheses appear as grey circles and the ground-truth as green stars. In accordance with our experiments, the spatial focus-of-attention recover most of the ground-truth elements. Though it adds an overhead, it is always better use the focus-of-attention than to run the sliding-box across the region S . Left column shows focus-of-attention build setting a low threshold in the inters-point detector. Right the focus-of-attention using a higher threshold.

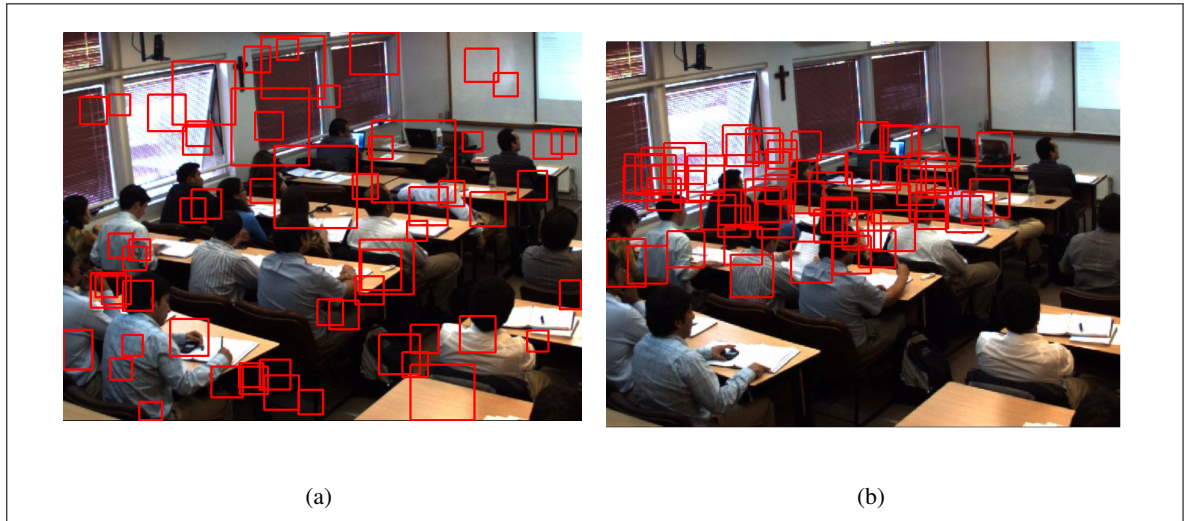


Figure 5.11. Random sample of 60 windows evaluated: (a) by a 2D detector, and (b) by our approach. In (a), windows must change their size in order to predict the real size of the object. We improve the search by using the real size of the object what limit to the size and locations of the projected windows, as shown in (b).

Image Example	Score	Image Example	Score	Image Example	Score
	-3.59		0.74		-2.20
	-2.58		1.24		-3.28
	-1.84		1.66		-4.54
	1.17		-1.92		-3.83
(a)		(b)		(c)	

Figure 5.12. Alignment test of collection of aspects. The figure shows the results of using sliding-box to evaluate among a set of hypotheses and their scores. Box scores are high when the box belongs to the head class, and its projections yield better alignment, as shown in (a) and (b). In (c) we observe background examples and their scores, all of which were negatives.

Chapter 6. DISCUSSION

This chapter presents general and specific remarks about our research, and summarizes the contribution of this thesis. We also discuss the implications of including information from others viewpoints in a multiple views system, the use of spatial focus-of-attention, and the methodology that we used for classifying and detecting people in multiple view indoor environments. Finally, we present future avenues of research following this approach.

6.1. Conclusions

This thesis describes our approach to detecting people in indoor multiple view environments. Our main hypothesis is that we can generate a set of potential candidates of a target object class using a corresponding projections of a parallelepiped defined in the 3D space and tailored to the size of the target object class, that we call *sliding-box*. This *sliding-box* passes through the three planes (X, Y, Z) of the spatial domain in the scene in which people are likely to appear. This approach allows us to search the target object class based on the physical dimensions and location of the this target in a scene. It also allows information from various viewpoints to be combined based on their corresponding regions, and finally let us to enhance the effectiveness of single view detection methods described in the state-of-the-art.

We focus our research on head detection in order to detect people in indoor environments. We take this assumption based on studies which claims that detecting the head helps avoid occlusions in crowded environments (Dalal & Triggs, 2005; Eshel & Moses, 2010; Ali & Dailey, 2012). Heads detection is an special case of people detection that presents an advantage because of heads are a less deformable objects than the entire body. In this way, heads detectors can be incorporated as part-based detector into more complex detections systems in order helps to find the complete body or the other body parts, and finally improve the detection (Zeng & Ma, 2010; Ali & Dailey, 2012; Xie et al., 2012; Nghiem et al., 2012; Chang et al., 2013; Hayashi et al., 2013).

We achieve the main objective of this research of designing and implementing a framework for heads detection based on a spatial version of the 2D sliding-windows technique. Our method extends this approach replacing the sliding-window with a *sliding-box*. In our implementation,

the detector runs this *sliding-box* across the 3D space through regions defined by a spatial focus-of-attention. Then, the algorithm generates projections of the *sliding-box* in each viewpoint. A multiple view classifier assigns a confidence value to each head hypothesis. Finally, a non-maximal suppression procedure helps to eliminate overlapped detection in order to obtain correct head detections.

To carry out the main goal of this thesis, we reach the three intermediate goals that we describe below:

- (i) We build a multiple view environment in order to emulate indoor conditions and generate multiple view datasets to train and test our method. These datasets allowed us to train multiple view models based on the information from collection of aspects, and evaluate performances in detection. We have published the datasets on the website <http://web.eng.puc.cl/~cppierin/projects> in order to contribute to other researches related to people detections using multiple views.
- (ii) Using these data, we design, develop and evaluate our framework for detecting people in indoor environments.
- (iii) We also extend the proposed approach to other detection problems. We specifically apply our approach to heads detection in the context of people detection. However, it can also be applied to other target objects classes and others problems that use multiple views, as we demonstrated with the detection of bubbles in materials (Pieringer & Mery, 2010). In this sense, our method would require adjust the *sliding-box* shape according to the target object classes, and the features that we use to represent them.

Single view detectors achieve good levels of detection performance, however, these algorithms are not always able to handle all target object classes in one camera because object variation or illumination changes. One of the most powerful contributions of our approach is the ability to integrate information coming from multiple view without using a matching technique. This advantage would allow us to improve detections provided by a single view detector in a multiple view environment by merging them into a single spatial detection. This integration makes it possible to recover loss detections generated by a 2D detector. This is an important issue in scenes in which people's position may change, and where we do not know where the detector will perform best in advance. We

used simultaneously all of the information in the scene at the same time. This means that we did not have to choose or train our approach on the best camera.

After experiments and testings, our approach yielded the four main proposed contributions:

- (i) a method to combine multiple view information based on a sliding box approach that allow us to reduce the correspondence problem at level of detection object, instead of using matching methods based on low-level pixel or interest points
- (ii) a framework that includes useful 3D cues and allow us to focus in the relevant parts of the 3D world in order to filter false candidates
- (iii) a guided search of the target object class by moving the sliding-box within a limited 3D space of interest.
- (iv) a classification approach based on combining information from multiple views that allows us to enrich the data used to train the models
- (v) and finally, the verification of the relevance of the previous ideas for the case of people counting using head detection showing a substantial increase in recognition performance with respect to alternative state-of-the-art techniques.

According to these contributions and our results, the main conclusions in relation to the topics presented in this thesis are detailed below:

- Detection results show that the one-vs-one (OVO-EC) ensemble of classifiers generally presents better detection performance in terms of average precision (AP). This difference in performance is more noticeable in test dataset *sq-01*, where heads present more pose variations. The performances of the three classification schemes were close in test dataset *sq-02*. In this case, the heads were turned to the front in some of the images. We conclude that the ensemble of classifiers approach allows us to represent each projection of the *sliding-box* in the collection of aspects using a compact representation. Therefore, we can imagine the first layer of the ensemble as a group of mid-level features. The second layer performs the mixture of these parts. Similar conclusions have been presented in (Sabzmejdani & Mori, 2007; Boureau et al., 2010), where authors proposed using mid-level features.

- The DPM detector based on epipolar geometry have the ability of discarding false detections in this multiple views configuration. However, this method is not always effective for associating detections that showed significant differences between frontal and rear views of the same head. By contrast, our *sliding-box* approach is able to code the structure inherent to the collection of aspects that is the object structure, the head structure in our case. In doing so, we are able to detect simultaneously using all of the cameras and at the same time and discarded an significant number of false positives generated as potential candidates.
- Results of heat-maps show the differences among our classification approaches. All our approaches achieve to detect correctly most of the ground-truth elements. However, we find that OVO-CE shows less spurious detections around the ground-truth neighborhood. The effectiveness of this procedure produces spatial hypotheses around the ground-truth and simultaneously it rejects large areas close to the edges of the region of interest S .
- Performance curves also show that main differences in detection rates are due to the suppression method (NMS) used to eliminate overlapped detections. Suppression based on the weighted Mean-Shit (WMS-NMS) improves the performance in both datasets. However, the quality of the results decreased in low detection thresholds in *sq-02*. We conclude that WMS-NMS is not able to correctly join overlapped detection at these low confidence values because of these extra noisy detections.
- The sensitive analysis of the focus-of-attention shows of that this attention mechanism helps to further restrict the search space and drastically reduce the number of head hypotheses. Though this attention mechanisms generate overhead, they pay off due to the complexity of detection (Papageorgiou & Poggio, 2000; Rutishauser et al., 2004; Frin-trop et al., 2005). Other single view methods (Viola & Jones, 2004; Felzenszwalb et al., 2010) and multiple-views approaches (Fleuret et al., 2008; Eshel & Moses, 2010) have already applied this idea. Figure 5.9 shows the ability of our focus-of-attention to retrieve candidate detections depends on the detection threshold of the interest-point detector. We show that is possible to choose a threshold in which we can recover all of the heads present in the scene. Table 5.3 shows that setting this threshold to -1.0 allow us to reduce the burden of the analysis into 85% in the case of *sq-01* and into 92% in *sq-02*. Setting the confidence threshold of the detector above to 0 we reduce drastically

the burden during detection, however, we report a significant decrease of the recall that may impact the overall performance of our detector.

- Although some remarkable single view approaches utilize spatial information (Salas & Tomasi, 2011; Spinello & Arras, 2011; Espinace et al., 2013), single view detection approaches generally do not take into account this information, and the inference of the objects size should be determined by estimating their scale. Our approach can recover this crucial information during the detection without using additional hardware. This spatial information was used for two purpose: to find the most likely location of the people; and to efficiently use the size of the objects. Therefore, it provides the detector with a more efficient search mechanism than a simple multi-scale estimation.
- On the one hand, calibration allows our method to use and recover the spatial information from the scene. It also allows us to combine information among viewpoints. Qualitative results of using geometrical cues to detection show that without a doubt calibration may be a useful tool in environments where we need to locate a target object classes in indoor environments with prohibited locations such as train stations, airports, etc. Detection results show that a simple mask for 2D detectors could not be enough to filter false detection. However, on the other hand calibration processes also is a limitation that makes our approach somewhat rigid to the scene structure.
- The pipeline of our method includes different types of processing that takes an average of 6 minutes to classify 1500 boxes in our MATLAB implementation, running on Ubuntu-Linux and a AMD Phenom II X4-925, 4GB RAM, 2.8GHz computer. This time depends on the total numbers of hypotheses generated after the focus-of-attention. Despite our promising detection results, processing time is a limitation. In the course of detection, the multiple view model evaluated each hypothesis considering its best alignment and chose the most confident collection of aspects. For now, this limitation restricts the use of our framework to small indoor spaces, similar to those we used in our experiments.

6.2. Future Work

Although our algorithm still does not yet work in real-time, we believe that our results open new path for future research on this subject. The same approach can be adapted to improve detections in other challenging scenarios using multiple views. Though the proposed approach offers also

advantages by incorporating information from multiple views, there is still room for improvements. The paragraphs below describe potential enhancements and extensions of the approach.

- There is no doubt that calibration is a required procedure so that the spatial information can be used and combined. However, this dependence with the environment makes difficult to ensure that calibration does not present changes affecting the projections stages. We believe that it is possible to extend our method to a semi-calibrated or an un-calibrated system, in which we do not require to use the calibration pattern for each new scene. There are techniques such as Structure From Motion (Szeliski, 2010; Agarwal et al., 2011) that allows recovering the 3D structure of a scene from 2D images without previous calibration. However, un-calibrated systems cannot recover metric information that we use in our method to define the geometric cues. One alternative may be using landmarks that helps to the structure estimation recover this spatial information and limit the calibration matrices computation to this measurements.
- Features are a relevant part of machine learning algorithms. Our method uses state-of-art features that have shown to represent successfully various target objects classes. Last representations of visual features use sparse codes algorithms to discover basis functions that captures high-level features in image data. This high-level features improve the classification and detection performance (Yang et al., 2009; C. Zhang et al., 2013). The integration of this algorithms to our method is an interesting avenue to explore.
- We train the classifiers using collection of aspect that we aligned manually. Then, during detection we search the best collection of aspect applying shifts over the collection define by each *sliding-box*. We believe that it is possible to improve the training and detection by including mechanisms for detecting latent parts and the head pose variations in order to learn the head structure automatically using classifiers that handle structure in data such as structural SVM or latent structural SVM (Tsochantaridis et al., 2004; Yu & Joachims, 2009).
- Currently, our approach requires that all cameras contain information to generate the collection of aspects. However, all of this data may not be present or may be noisy during occlusion situations. In this case, our method can be improved using a more flexible model that can handle this missing data.

- Our research focus on head detection in still images of multiple view environments. In this sense, incorporating a tracking strategy would allow us to estimate trajectories of these heads in 3D spaces. We believe that the use of dynamic information, can improve detection in surveillance systems.

References

- Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S. M., et al. (2011). Building Rome in a day. *Communications of the ACM*, 54(10), 105–112.
- Aghajan, H., & Cavallaro, A. (2009). *Multi-camera networks: principles and applications*. Academic press.
- Alahi, A., Ortiz, R., & Vandergheynst, P. (2012). FREAK: Fast Retina Keypoint. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CPVR)* (pp. 510–517). IEEE.
- Alexe, B., Deselaers, T., & Ferrari, V. (2012). Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(11), 2189–2202.
- Ali, I., & Dailey, M. (2012). Multiple human tracking in high-density crowds. *Image and Vision Computing*, 30(12), 966–977.
- Ali Shah, S., Bennamoun, M., Boussaid, F., & El-Sallam, A. A. (2013). Automatic object detection using objectness measure. In *Proceedings of the International Conference on Communications, Signal Processing, and their Applications (ICCSPA)* (pp. 1–6).
- Anand, R., Kumar, P., et al. (2006). Flaw detection in radiographic weld images using morphological approach. *NDT & E International*, 39(1), 29–33.
- Atmosukarto, I., & Shapiro, L. G. (2013). 3D object retrieval using salient views. *International Journal of Multimedia Information Retrieval*, 2(2), 103–115.
- Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3), 346–359.

- Benenson, R., Timofte, R., & Van Gool, L. (2011). Stixels estimation without depth map computation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 2010–2017).
- Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1), 1–127.
- Bishop, C., et al. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer New York.
- Boerner, H., & Strecker, H. (1988). Automated x-ray inspection of aluminum castings. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 10(1), 79–91.
- Borji, A., Sihite, D. N., & Itti, L. (2012). Salient object detection: A benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 414–429). Springer.
- Bosch, A., Zisserman, A., & Muñoz, X. (2007). Image classification using random forests and ferns. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1–8).
- Boureau, Y.-L., Bach, F., LeCun, Y., & Ponce, J. (2010). Learning mid-level features for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2559–2566).
- Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123–140.
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.
- Bustos, B., Keim, D. A., Saupe, D., Schreck, T., & Vranić, D. V. (2005). Feature-based similarity search in 3D object databases. *ACM Computing Surveys*, 37, 345–387.

- Calonder, M., Lepetit, V., Strecha, C., & Fua, P. (2010). Brief: Binary robust independent elementary features. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 778–792). Springer.
- Carrasco, M., & Mery, D. (2006). Automated visual inspection using trifocal analysis in an uncalibrated sequence of images. *Materials Evaluation*, 64(9), 900–906.
- Chang, R., Chua, T. W., Leman, K., Wang, H. L., & Zhang, J. (2013). Automatic cooperative camera system for real-time bag detection in visual surveillance. In *Proceedings of the International Conference on Distributed Smart Cameras (ICDSC)* (pp. 1–6).
- Cheng, M.-M., Zhang, Z., Lin, W.-Y., & Torr, P. (2014). BING: Binarized normed gradients for objectness estimation at 300fps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3286–3293).
- Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 17(8), 790–799.
- Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(5), 603–619.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.
- Cristani, M., Farenzena, M., Bloisi, D., & Murino, V. (2010). Background subtraction for automated multisensor surveillance: a comprehensive review. *EURASIP Journal on Advances in signal Processing*, 2010, 43.
- Cyr, C. M., & Kimia, B. B. (2001). 3D Object recognition using shape similarity-based aspect graph. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Vol. 1, pp. 254–261).

Cyr, C. M., & Kimia, B. B. (2004). A similarity-based aspect-graph approach to 3D object recognition. *International Journal of Computer Vision*, 57(1), 5–22.

Dalal, N. (2006). *Finding people in images and videos*. Unpublished doctoral dissertation, Institut National Polytechnique de Grenoble-INPG.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 1, pp. 886–893).

Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 428–441). Springer.

Davis, J. W., Morison, A. M., & Woods, D. D. (2007). An adaptive focus-of-attention model for video surveillance and monitoring. *Machine Vision and Applications*, 18(1), 41–64.

Dean, T., Ruzon, M. A., Segal, M., Shlens, J., Vijayanarasimhan, S., & Yagnik, J. (2013). Fast, accurate detection of 100,000 object classes on a single machine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1814–1821).

Dee, H., & Velastin, S. (2007). How close are we to solving the problem of automated visual surveillance? *Machine Vision and Applications*, 1–15.

Delannay, D., Danhier, N., & De Vleeschouwer, C. (2009). Detection and recognition of sports (wo) men from multiple views. In *Proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)* (pp. 1–7).

Dollar, P., Appel, R., Belongie, S., & Perona, P. (2014). Fast feature pyramids for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36(8), 1532–1545.

- Dollar, P., Tu, Z., Perona, P., & Belongie, S. (2009). Integral Channel Features. In *British Machine Vision Conference (BMVC)* (Vol. 2, p. 5).
- Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: an evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1–20.
- Eshel, R., & Moses, Y. (2010). Tracking in a dense crowd using multiple cameras. *International Journal of Computer Vision*, 88(1), 129–143.
- Espinace, P., Kollar, T., Roy, N., & Soto, A. (2013). Indoor scene recognition by a mobile robot through adaptive object detection. *Robotics and Autonomous Systems*, 61(9), 932–947.
- Evans, M., Osborne, C. J., & Ferryman, J. (2013). Multicamera object detection and tracking with object size estimation. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 177–182).
- Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2), 303–338.
- Everingham, M., Zisserman, A., Williams, C., & Van Gool, L. (2006). The PASCAL Visual Object Classes Challenge 2006 (VOC2006) results.
- Farhadi, A., & Tabrizi, M. K. (2008). Learning to recognize activities from the wrong view point. In *Computer vision—eccv 2008* (pp. 154–166). Springer.
- Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 2, pp. 524–531).

- Felzenszwalb, P., Girshick, R., & McAllester, D. (2010). Cascade object detection with deformable part models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2241–2248).
- Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2009). Object Detection with Discriminatively Trained Part Based Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 99.
- Ferrari, V., Tuytelaars, T., & Van Gool, L. (2006). Simultaneous object recognition and segmentation from single or multiple model views. *International Journal of Computer Vision*, 67(2), 159–188.
- Ferreira, A., Marini, S., Attene, M., Fonseca, M., Spagnuolo, M., Jorge, J., et al. (2010). Thesaurus-based 3D object retrieval with part-in-whole matching. *International Journal of Computer Vision*, 89, 327–347.
- Fleuret, F., Berclaz, J., Lengagne, R., & Fua, P. (2008). Multi-camera people tracking with a probabilistic occupancy map. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(2), 267–282.
- Freund, Y., & Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *Proceedings of the European Conference on Computational Learning Theory* (pp. 23–37).
- Frintrop, S., Backer, G., & Rome, E. (2005). Goal-directed search with a top-down modulated computational attention system. *Pattern Recognition*, 117–124.
- Frintrop, S., Rome, E., & Christensen, H. I. (2010). Computational visual attention systems and their cognitive foundations: A survey. *ACM Transactions on Applied Perception (TAP)*, 7(1), 6.

- Fukunaga, K., & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1), 32–40.
- Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Han, B., Joo, S.-W., & Davis, L. S. (2011). Multi-camera tracking with adaptive resource allocation. *International Journal of Computer Vision*, 91(1), 45–58.
- Hartley, R., & Zisserman, A. (2003). *Multiple View Geometry in Computer Vision* (Vol. 2). Cambridge University Press.
- Hayashi, M., Yamamoto, T., Aoki, Y., Ohshima, K., & Tanabiki, M. (2013). Head and upper body pose estimation in team sport videos. In *Proceedings of the IAPR Asian Conference on Pattern Recognition (ACPR)* (pp. 754–759).
- Heikkila, J., & Silven, O. (1997). A four-step camera calibration procedure with implicit image correction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (p. 1106-1112).
- Helmer, S., & Lowe, D. (2010). Using stereo for object recognition. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3121–3127).
- Hetzel, G., Leibe, B., Levi, P., & Schiele, B. (2001). 3D object recognition from range images using local feature histograms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 2, p. II-394-II-399 vol.2).

- Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).
- Hu, W., Tan, T., Wang, L., & Maybank, S. (2004). A survey on visual surveillance of object motion and behaviors. In *IEEE Transactions on Systems, Man and Cybernetics, Part C*. (Vol. 34, pp. 334–352).
- Ikeuchi, K., & Kanade, T. (1988). Automatic generation of object recognition programs. *Proceedings of the IEEE*, 76(8), 1016–1035.
- Itti, L. (2000). *Models of bottom-up and top-down visual attention*. Unpublished doctoral dissertation, California Institute of Technology.
- Itti, L., & Koch, C. (1999). Learning to detect salient objects in natural scenes using visual attention. In *Image understanding workshop* (pp. 1201–1206).
- Kadir, T., Zisserman, A., & Brady, M. (2004). An affine invariant salient region detector. In (Vol. 1, pp. 228–241). Springer.
- Kazhdan, M., Funkhouser, T., & Rusinkiewicz, S. (2003). Rotation invariant spherical harmonic representation of 3d shape descriptors. In *Proceedings of the Eurographics Symposium on Geometry processing (SIGGRAPH)* (pp. 156–164).
- Khan, S., & Shah, M. (2006). A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *Proceedings of the European Conference on Computer Vision (ECCV)* (Vol. 3954, p. 133-146).
- Kim, K., Chalidabhongse, T., Harwood, D., & Davis, L. (2005). Real-time foreground-background segmentation using codebook model. *Real-time Imaging*, 11(3), 172–185.

- Kim, K., & Davis, L. (2006). Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 98–109). Springer.
- Koenderink, J. J., & Doorn, A. J. van. (1979). The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32(4), 211–216.
- Kuno, Y., Watanabe, T., Shimosakoda, Y., & Nakagawa, S. (1996). Automated detection of human for visual surveillance system. In *Proceedings of the International Conference on Pattern Recognition (CVPR)* (Vol. 3, pp. 865–869).
- Kushal, A., Schmid, C., & Ponce, J. (2007). Flexible object models for category-level 3D object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).
- Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 2, pp. 2169–2178).
- Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 609–616).
- Leutenegger, S., Chli, M., & Siegwart, R. Y. (2011). BRISK: Binary Robust invariant scalable keypoints. In *Proceedings of the International Conference on Computer Vision (ICCV)* (pp. 2548–2555). IEEE.
- Li, X., Guskov, I., & Barhak, J. (2006). Robust alignment of multi-view range data to CAD model. In *IEEE International Conference on Shape Modeling and Applications, 2006 (SMI 2006)*. (pp. 17–17).

Liem, M. C., & Gavrilu, D. M. (2013). A comparative study on multi-person tracking using overlapping cameras. In *Computer Vision Systems* (pp. 203–212). Springer.

Liem, M. C., & Gavrilu, D. M. (2014). Joint multi-person detection and tracking from overlapping cameras. *Computer Vision and Image Understanding*.

Lindeberg, T. (1993). *Scale-space theory in computer vision*. Kluwer Academic Print on Demand.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., et al. (2011). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(2), 353–367.

Lo, B., & Velastin, S. (2001). Automatic congestion detection system for underground platforms. In *Proceedings of the International Symposium on Intelligent Multimedia, Video and Speech* (pp. 158–161).

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Maji, S., Berg, A. C., & Malik, J. (2008). Classification using intersection kernel support vector machines is efficient. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).

Marchesotti, L., Cifarelli, C., & Csurka, G. (2009). A framework for visual saliency detection with applications to image thumbnailing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2232–2239).

Matas, J., Chum, O., Urban, M., & Pajdla, T. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10), 761–767.

- Mery, D., Filbert, D., & Jaeger, T. (2005). Image Processing for Fault Detection in Aluminum Castings. In *Analytical Characterization of Aluminum and Its Alloys* (pp. 701–738). Ed. C.S. MacKenzie and G.E. Totten, CRC Press, Taylor and Francis, Florida.
- Mery, D., Jaeger, T., & Filbert, D. (2002). A review of methods for automated recognition of casting defects. *Insight-Non-Destructive Testing and Condition Monitoring*, 44(7), 428–436.
- Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1), 63–86.
- Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27(10), 1615–1630.
- Mittal, A., & Davis, L. (2003). M2 Tracker: A multi-view approach to segmenting and tracking people in a cluttered scene. *International Journal of Computer Vision*, 51(3), 189–203.
- Montabone, S., & Soto, A. (2010). Human Detection Using a Mobile Platform and Novel Features Derived From a Visual Saliency Mechanism. *Image and Vision Computing*, 28(3), 391–402.
- Mu, Y., Yan, S., Liu, Y., Huang, T., & Zhou, B. (2008). Discriminative local binary patterns for human detection in personal album. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).
- Mundy, J. L. (2006). Object recognition in the geometric era: A retrospective. In *Toward Category-level Object Recognition* (pp. 3–28). Springer.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nghiem, A. T., Auvinet, E., & Meunier, J. (2012). Head detection using kinect camera and its application to fall detection. In *International Conference on Information Science, Signal Processing and their Applications (ISSPA)* (pp. 164–169).

Nixon, M., & Aguado, A. (2012). *Feature Extraction & Image Processing for Computer Vision* (3rd ed.). Academic Press.

Ojala, T., Pietikäinen, M., & Mäenpää, T. (2000). Gray scale and rotation invariant texture classification with local binary patterns. In *Proceedings of the European Conference on Computer Vision (ECCV)* (Vol. 1842, p. 404-420). Springer.

Papageorgiou, C., & Poggio, T. (2000). A trainable system for object detection. *International Journal of Computer Vision*, 38(1), 15–33.

Park, D., Ramanan, D., & Fowlkes, C. (2010). Multiresolution models for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 241–254). Springer.

Pedersoli, M., González, J., Bagdanov, A., & Villanueva, J. (2010). Recursive coarse-to-fine localization for fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 280–293). Springer.

Pedersoli, M., Gonzalez, J., Hu, X., & Roca, X. (2014). Toward real-time pedestrian detection based on a deformable template model. *IEEE Transactions on Intelligent Transportation Systems*, 15(1), 355-364.

Pei, W.-J., Zhang, Y.-L., Zhang, Y., & Zheng, C.-H. (2014). Pedestrian detection based on hog and lbp. In *Intelligent computing theory* (pp. 715–720). Springer.

- Pieringer, C., & Mery, D. (2010). Flaw detection in aluminium die castings using simultaneous combination of multiple views. *Insight-Non-Destructive Testing and Condition Monitoring*, 52(10), 548–552.
- Pieringer, C., Mery, D., & Soto, A. (2012). Head modeling using multiple-views. In *Proceedings of the Chilean Workshop on Pattern Recognition (CWPR)* (pp. 40–43).
- Platt, J. C. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*.
- Poggio, T., & Edelman, S. (1990). A network that learns to recognize 3D objects. *Nature*, 343(6255), 263–266.
- Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3), 21–45.
- Pontil, M., & Verri, A. (1998). Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(6), 637–646.
- Porikli, F., & Tuzel, O. (2006). Fast construction of covariance matrices for arbitrary size image windows. In *Proceedings of the IEEE International Conference on Image Processing* (pp. 1581–1584).
- Rahtu, E., Kannala, J., & Blaschko, M. (2011). Learning a category independent object detection cascade. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1052–1059).
- Rosten, E., & Drummond, T. (2006). Machine learning for high-speed corner detection. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 430–443). Springer.

- Rothganger, F., Lazebnik, S., Schmid, C., & Ponce, J. (2006). 3D Object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, 66(3), 231–259.
- Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: an efficient alternative to SIFT or SURF. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 2564–2571).
- Rutishauser, U., Walther, D., Koch, C., & Perona, P. (2004). Is bottom-up attention useful for object recognition? In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. (Vol. 2, pp. II–37).
- Sabzmeydani, P., & Mori, G. (2007). Detecting pedestrians by learning shapelet features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1–8).
- Salas, J., & Tomasi, C. (2011). People detection using color and depth images. In *Pattern recognition* (pp. 127–135). Springer.
- Savarese, S., & Fei-Fei, L. (2007). 3D generic object categorization, localization and pose estimation. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1–8.
- Savarese, S., & Fei-Fei, L. (2010). Multi-view object categorization and pose estimation. *Computer Vision*, 205–231.
- Schapire, R. (1990). The strength of weak learnability. *Machine learning*, 5(2), 197–227.
- Schwartz, W. R., Kembhavi, A., Harwood, D., & Davis, L. S. (2009). Human detection using partial least squares analysis. In *Proceedings of the International Conference on Computer Vision (ICCV)* (pp. 24–31).

Shen, X., & Wu, Y. (2012). A unified approach to salient object detection via low rank matrix recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 853–860).

Shuai, B., Cheng, Y., Li, S., & Su, S. (2012). A hierarchical clustering based non-maximum suppression method in pedestrian detection. In *Proceedings of the Intelligent Science and Intelligent Data Engineering* (pp. 201–209). Springer.

Siva, P., Russell, C., Xiang, T., & Agapito, L. (2013). Looking beyond the image: Unsupervised learning for object saliency and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3238–3245).

Song, M., Tao, D., & Maybank, S. J. (2013). Sparse camera network for visual surveillance – a comprehensive survey. *arXiv preprint arXiv:1302.0446*.

Song, Z., & Klette, R. (2013). Robustness of Point Feature Detection. *Computer Analysis of Images and Patterns*, 91–99.

Spinello, L., & Arras, K. O. (2011). People detection in rgb-d data. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3838–3843).

Stauffer, C., & Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (Vol. 2).

Stone, J. V. (1999). Object recognition: View-specificity and motion-specificity. *Vision Research*, 39(24), 4032–4044.

Su, H., Sun, M., Fei-Fei, L., & Savarese, S. (2009). Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *International Conference on Computer Vision 2009 (ICCV2009)*.

Su, T.-M., Lin, C.-C., Lin, P.-C., & Hu, J.-S. (2006). Shape memorization and recognition of 3d objects using a similarity-based aspect-graph approach. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Vol. 6, pp. 4920–4925).

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer.

Tarr, M. J., & Kriegman, D. J. (2001). What defines a view? *Vision Research*, 41(15), 1981–2004.

Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., Schiele, B., & Van Gool, L. (2006). Towards multi-view object class detection. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 1589–1596).

Torralba, A., Murphy, K. P., & Freeman, W. T. (2004). Sharing features: efficient boosting procedures for multiclass object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 2, pp. II–762).

Toshev, A., Makadia, A., & Daniilidis, K. (n.d.). Shape-based object recognition in videos using 3D synthetic object models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages=288–295, year=2009, organization=IEEE.

Tsochantaridis, I., Hofmann, T., Joachims, T., & Altun, Y. (2004). Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the International Conference on Machine Learning (ICML)* (p. 104).

Tuzel, O., Porikli, F., & Meer, P. (2006). Region covariance: A fast descriptor for detection and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 589–600). Springer.

Tuzel, O., Porikli, F., & Meer, P. (2008). Pedestrian detection via classification on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(10), 1713–1727.

Ulrich, M., Wiedemann, C., & Steger, C. (2012). Combining scale-space and similarity-based aspect graphs for fast 3d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10), 1902–1914.

Valera, M., & Velastin, S. (2005). Intelligent distributed surveillance systems: a review. In *IEEE Proceedings on Vision, Image and Signal Processing* (Vol. 152, pp. 192–204).

Vedaldi, A., & Fulkerson, B. (2010). Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia* (pp. 1469–1472).

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Vol. 1, pp. I–511).

Viola, P., & Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2), 137–154.

Viola, P., Jones, M., & Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2), 153–161.

Wang, X., Han, T., & Yan, S. (2009). An HOG-LBP human detector with partial occlusion handling. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 32–39).

- Wren, C. R., Azarbayejani, A., Darrell, T., & Pentland, A. P. (1997). Pfnder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7), 780–785.
- Wu, S., & Lew, M. (2013). Evaluation of salient point methods. In *Proceedings of the ACM International Conference on Multimedia (MM)* (pp. 685–688). ACM Press.
- Xie, D., Dang, L., & Tong, R. (2012). Video based head detection and tracking surveillance system. In *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* (pp. 2832–2836).
- Yang, J., Yu, K., Gong, Y., & Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1794–1801).
- Yildiz, A., & Akgul, Y. S. (2012). A fast method for tracking people with multiple cameras. In *Trends and topics in computer vision* (pp. 128–138). Springer.
- Yu, C.-N. J., & Joachims, T. (2009). Learning structural SVMs with latent variables. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 1169–1176).
- Zaytseva, E., & Vitria, J. (2012). A search based approach to non maximum suppression in face detection. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)* (pp. 1469–1472).
- Zeng, C., & Ma, H. (2010). Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. In *Proceedings of the International Conference on Pattern Recognition (ICPR)* (pp. 2069–2072).
- Zhang, C., Wang, S., Huang, Q., Liu, J., Liang, C., & Tian, Q. (2013). Image classification using spatial pyramid robust sparse coding. *Pattern Recognition Letters*, 34(9), 1046–1052.

Zhang, Z. (1999). Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Vol. 1, pp. 666–673).

Zia, M. Z., Stark, M., Schiele, B., & Schindler, K. (2013). Detailed 3D representations for object recognition and modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11), 2608–2623.

APPENDICES

APPENDIX A. FLAW DETECTIONS IN ALUMINIUM DIE CASTING

This appendix presents the concept testing of using our approach to solve a simple and known detection problem. We summarize the results of this test in Pieringer and Mery (2010), which includes evaluation of the projection method and an the ensemble scheme to classify flaws in aluminum die casting in multiple views. Defects detection is a relevant problem in manufacture industry. Although human inspectors are hard to replace, they can fulfill these requirements only for short periods. We conclude two important contributions: simulated flaws can be effectively used to train classifiers used in these applications, due to real flaws are rare events in industrial manufacturing processes; and simultaneous combination of information from different points of view using sliding-boxes is a robust approach to flaw identification.

Flaw Detection in Aluminum Die Castings Using Simultaneous Combination of Multiple Views

Christian Pieringer
cppierin@uc.cl

Domingo Mery
dmery@ing.puc.cl

Álvaro Soto
asoto@ing.puc.cl

Abstract

Recently, X-rays have been adopted as the principal non-destructive testing method to identify flaws within an object which are undetectable to the naked eye. Automatic inspection using radiographic images has been made possible by incorporating image processing techniques into the process. In a previous work, we proposed a framework to detect flaws in aluminum castings using multiple views. The process consisted of flaw segmentation, matching, and finally tracking the flaws along the image sequence. While the previous approach required effective segmentation and matching algorithms, this investigation focuses on a new detection approach. The proposed method combines, simultaneously, information gathered from multiple views of the scene, this does not require searching for correspondences or matching. By gathering all the projections from a 3D point, obtained from a sliding box in the 3D space, we train a classifier to learn to detect simulated flaws using all the evidence available. This paper describes our proposed method and presents its performance record in flaw detections using various classifiers. Our approach yields promising results, 94% of true positives detected with 95% sensitivity in real flaws. We conclude that simultaneously combining information from different points of view is a robust approach to flaw identification.

Keywords — Automated inspection, flaw detection, saliency, computer vision, multiple views.

1 Introduction

Radioscopy has been embraced as the best tool for non-destructive testing (NDT) in industrial production given that most defects are not visible on the object's surface (Mery and Filbert, 2002). The material defects which occur during the casting process must be detected in order to satisfy safety requirements, consequently it is necessary to check 100% of the parts. Even though X-rays detect flaws in cast pieces, they often manifest as small and low contrast objects which are difficult to detect as seen in Fig. 1(a). Due to these difficulties it is necessary to incorporate image processing techniques that accurately high-

light flaws, while separating them from the background, then finally classifying the flaws correctly.

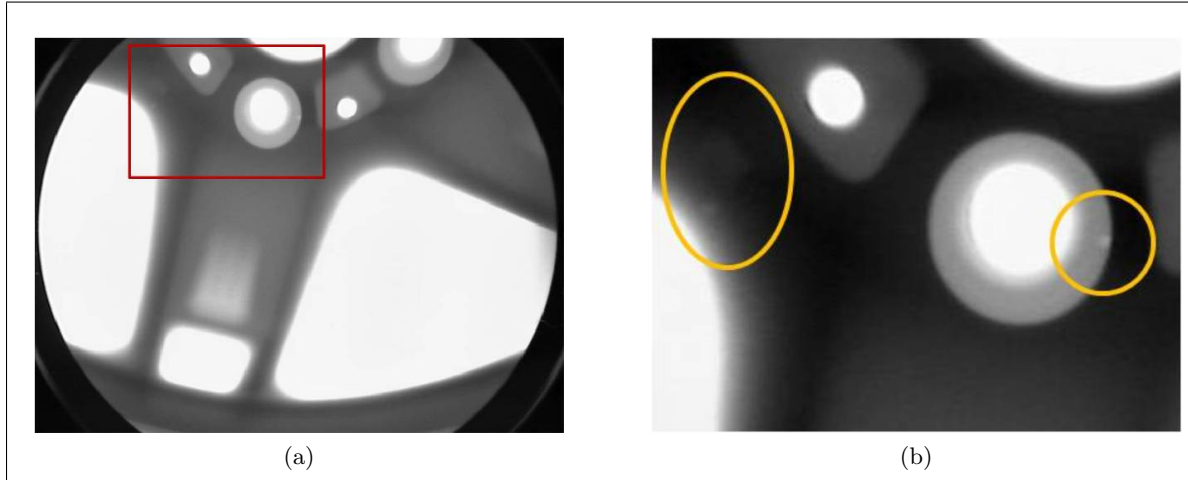


Figure 1: Flaw example in an aluminum wheel. (a) X-Ray image of aluminum wheel. (b) Magnification of the flaw areas. Circles denote a flaw within the piece.

A typical automated X-ray system is schematically presented and described in Mery and Filbert (2002). The process is generally performed in five steps:

- The *manipulator* places the casting in the desired position.
- The *X-ray tube* generates X-rays which pass through the casting.
- The X-rays are detected by a fluorescent entrance screen in the *image intensifier*, amplified and depicted onto a phosphor screen.
- The image intensifier converts the X-rays to a visible radiosopic image.
- The guided and focused image is registered by the *CCD-camera*. The *image processor* converts the analog video signal, transferred by the CCD-camera, into a digital data stream. Digital image processing is used to improve and evaluate the radiosopic image.

New X-Ray techniques utilize flat amorphous silicon detectors as image sensors in industrial inspection systems (Purschke, 2002). These detectors use a semi-conductor to convert energy from the X-ray into an electrical signal without an image intensifier. However due to their high cost, NDT using flat detectors is not as feasible as the use of image intensifiers. Various approaches to automated flaw detection in aluminum castings can be found in (Boerner and Strecker, 1988; Mery and Filbert, 2002; Mery et al., 2002a,b; Pizarro et al., 2008). Those works proposed flaw detection methods without obtaining a priori information of flaws or the object's structure. They developed methods to manage the lack of a priori

information because in real processes flaws are rare, making it is extremely difficult to get samples. However those methods are mainly supported in the pre-processing of the obtained images meaning they are dependent on the parameters of the processing algorithms and therefore inflexible to the variations of image intensity. This is where our research differs, we explore a new method of training classifiers with simulated samples of flaws which are easier to obtain than real samples, and then proceed to test the system by applying real flaws (Mery et al., 2005).

Multiple views drastically improve a systems detection rate by eliminating numerous false alarms, mainly because they provide additional or complementary information about the object being tested. This process involves using a sequence of X-ray images of a casting, all of which are taken from different positions. The next step is the segmentation of a flaw in one view followed by its subsequent tracking throughout the sequence. This approach can be applied to either calibrated or non calibrated sequences (Carrasco and Mery, 2006; Mery and Filbert, 2002).

Despite their advantages, the aforementioned methods still require effective segmentation in the first step in order to generate possible flaws to be tracked in following views. In addition, methods based on geometrical constraints to track the hypothetical flaws along the sequence, such as epipolar or trifocal tensors, require robust matching algorithms throughout different views. Motivated to eliminate these disadvantages, we investigate a new approach to combine information from multiple-views which allows for flexible learning without requiring a priori information of the object’s structure.

We propose using a sliding box, which moves within 3D space occupied by the casting object, to gather all projections from multiple views of this local space, Fig. 2(a). Our approach allows us to avoid matching because all the projections from the box are locally corresponding. In the following step we create a coarse detection of each flaw from the projections by using a saliency detector as in Achanta et al. (2008). Each detection is represented by two kinds of feature descriptors as proposed in Bosch et al. (2007): one for shape and one for appearance. Finally, we combine the responses of individual classifiers from all views which results in a final classification of each flaw sequence.

This paper is organized as follows. Section 2 introduces details of our approach, such as the searching method within the 3D space, the features used to describe aluminum casting flaws, as well as our methodology for combining information from multiple views. Sections 3 and 4 discuss our dataset and results, respectively. Finally, Section 5 presents conclusions about our approach.

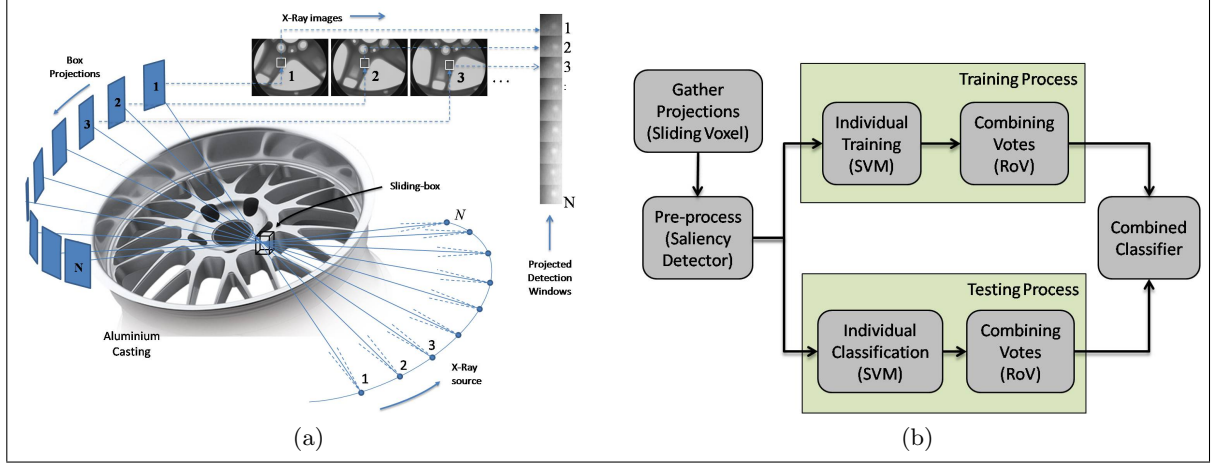


Figure 2: Proposed Methodology. (a) Shows process for gathering information from multiple views. (b) Process diagram of the proposed method.

2 Proposed Method

In a previous work, we proposed a framework for tracking flaws in castings (Mery and Filbert, 2002). This method has three main parts: flaw segmentation, matching of candidates in different views, and finally tracking of flaws in the image sequence. The system eliminates the flaws that are not tracked in third and fourth views.

Our method is based on the principal that it is possible to get results similar to previous works described in the introduction by:

1. replacing the segmentation stage with a more flexible approach based on object detection and recognition, and
2. replacing the matching and tracking stages with a complete and simultaneous analysis of the scene.

An automated flaw detection system first requires samples of the object being tested in order to train a classifier. Once this classifier is fully trained, it is used to detect previously unseen flaws, Fig 2(b). In the next four subsections we explain the main parts of our algorithm.

2.1 The Sliding Box

The state-of-the-art approach to machine learning, for object detection and recognition in single images, consists of running a sliding window over an image at different scales and finally detecting the class of interest within the image (Dalai et al., 2005; Viola and Jones, 2004).

We propose a slightly different approach to integrate information, received from a sliding box in a multiple view scheme, than methods which use epipolar geometry to establish



Figure 3: Flaw sequence constructed with information from multiple views. The display range was extended to facilitate viewing the results

correspondence and use matching as in Carrasco and Mery (2006); Mery and Filbert (2002); Mery et al. (2002a). A sliding box is used to scan the 3D space of the scene. Every time the box changes its position, projections from various points of view are simultaneously gathered. Once received, we implement a method to effectively combine the relevant information from multiple views.

Our automatic inspection method uses a calibrated X-Ray system, allowing us to project all 3D points within their corresponding view (Hartley and Zisserman, 2003). In essence, as the box moves through the space occupied by the casting, we project its vertices over each available view in the system. In the following step, we select the regions from each view that encompass the most projected points. Finally, each projected region which is considered valid are grouped together and arranged into a sequence that shows the flaw from all available views, as shown in Fig. 3.

2.2 Image Pre-Processing and Saliency Detector

The proposed method, based on object classification, requires the training of a classifier with flaw samples. However, this is difficult because flaws are very rare occurrences within images. They are generally difficult to detect because they appear within the image as a small section of low contrast pixels, even after being isolated in the flaw sequence. Before training the classifier it is necessary to obtain an extremely detailed sample. To do this, we extract the flaw's structure and remove the majority of background details.

At this stage we apply a salient detector to the inverse image of the selected projection. This detector doesn't use parameters, in fact it can adapt automatically to varying conditions of contrast in the samples. The results are coarser than the segmentation used in Mery and Filbert (2002) and Mery et al. (2002a), but it is still suitable for our approach due to its freedom of parameters. The effects of applying the detector over the original image and inverse image are different, see Fig.4(c) and Fig.4(f). Even though both results deliver salient zones, those obtained from the inverse image allow us to correctly and precisely isolate the flaw. Our saliency detector is based on the visual attention systems proposed in Achanta et al. (2008); Montabone and Soto (2009). Visual attention systems are inspired by primate visual systems, and accordingly we believe they better emulate inspections made by humans in flaw classification tasks. Saliency detection consists of computing local contrast between a specific region within an image and its surroundings using one or more features such as color, intensity and orientation. Thus the saliency of a region is high when its properties

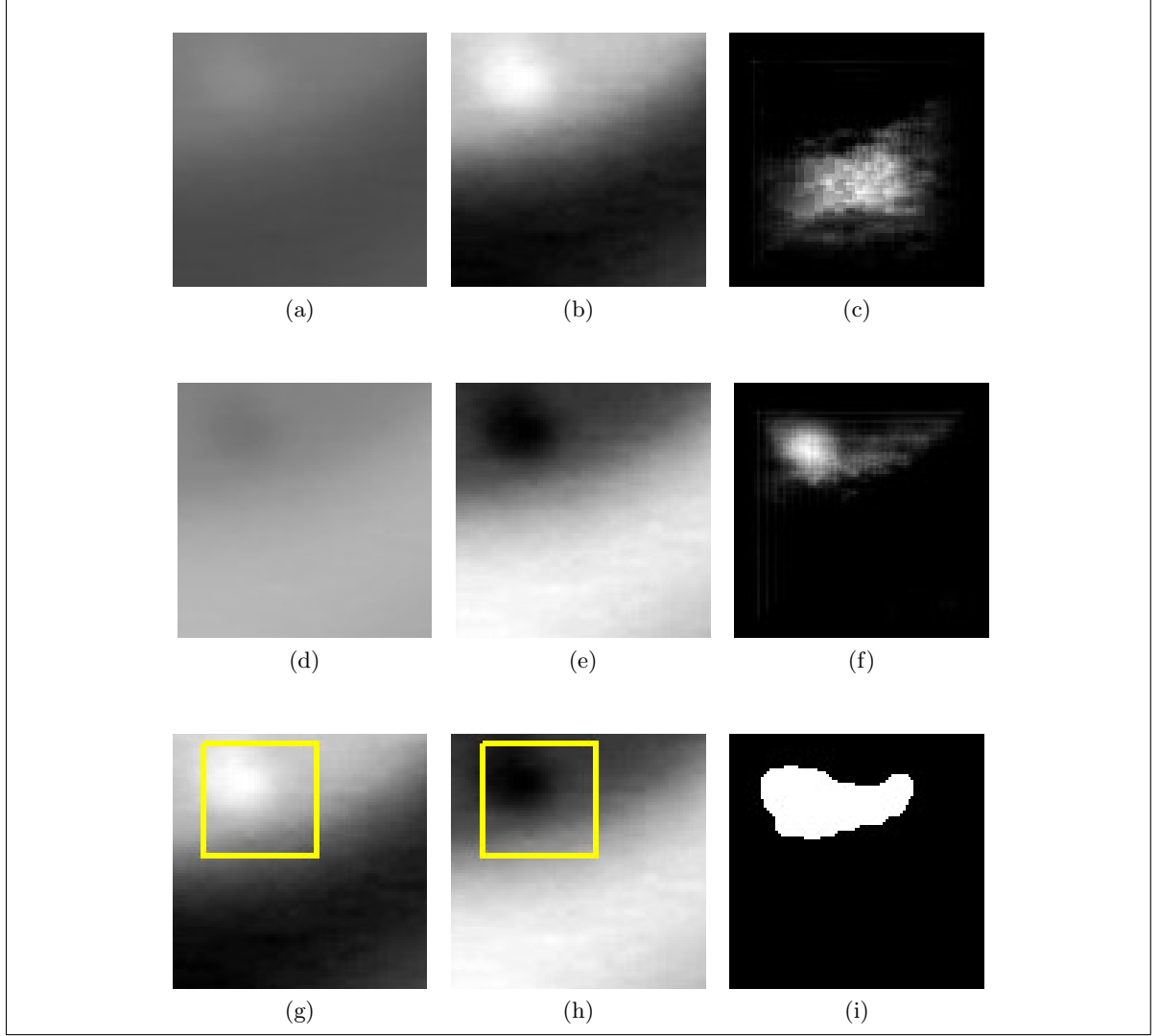


Figure 4: Results of saliency detector applied to original image and its complement. (a) Patch of the original flaw. (d) Complement version of the patch. (b) and (e) Extended display range of the original and complement patches, respectively. (c) and (f) Saliency maps.

differ considerably from the rest of the image (Achanta et al., 2008).

The detection process begins by obtaining the inverse images of the grey scale patches from each projection, as shown in Fig.4(d). Then, we apply the saliency detector to the modified patches. This results in a saliency map, see Fig.4(c) and Fig.4(f). We then determine a threshold to produce a binary image to remove noise from the edges of the saliency map, Fig.4(i). Finally, we choose a Region of Interest (ROI) around the maximum value of the saliency map Fig.4(g) and Fig.4(h). By trial and error we determined that the ROI served us best when set as 40 pixels. Projections without detections are classified as non flaws.

2.3 Features

Once we have an accurate flaw detected, after applying the saliency detector, we are able to extract three sets of features from the samples: Crossing Line Profile (CLP) (Mery, 2003), Pyramidal Histograms of Oriented Gradients (PHOG) (Bosch et al., 2007), and Histograms of Oriented Gradients based on SIFT descriptor (Lowe, 2004).

The selected features are used to represent both shape and appearance as in Bosch et al. (2007). The features vector for shape was constructed by concatenating features of CLP and PHOG. These features retrieve the circular shape of the flaws. The dimensions of feature vectors consist of 350 for shape and 128 for appearance.

CLP is defined as the best grey level profile along the length of a straight line within a region of interest (ROI). Eight profiles, distributed every $\pi/8$, are calculated. Each profile is normalized with respect to its average and standard deviation. Fourier Transform is extracted at the best profile and its Fourier coefficients are considered as features vector.

CLP has been proven to effectively represent circular flaws in radioscopic images. We combine the shape features vector and the vector with the best profile inside the bounding box with the flaw detection in the saliency image.

PHOG features have been successfully utilized in investigations dealing with recognition and classification of objects. These features use histograms to encode the gradient information in regards to the defined border of the object at different pyramidal levels. They perform better than the Chamfer Distance, which performs a template matching based on distance transform, because PHOG deals better with rotated images, it is an appropriate compact vector for learning with kernel based algorithms, and is flexible in regards to spatial correspondence. These features were calculated on the saliency map generated by the detector explained in 2.2, using only two levels of pyramids to avoid over fitting as suggested in Bosch et al. (2007), see Fig.5.

A SIFT descriptor computes locally and projects onto an image the histogram of oriented gradient over a point of interest at a given scale, orientation and position. In our method, we compute the gradient of oriented histograms as local descriptors over the detection bounding box resized to 16x16 pixels with neither scale nor orientation information since that information is already coded in the 3D model.

2.4 Classification and Combining Multiple-Views

Our classification process consists of two stages: individual classification and joint evaluation utilizing all available views.

First, every element in the flaw sequence as in Fig.3 was classified individually as a flaw or non-flaw. For this task, we trained two Support Vector Machines (SVMs) with polynomial kernels because SVMs can better manage large feature vectors, as in Bosch et al. (2007).

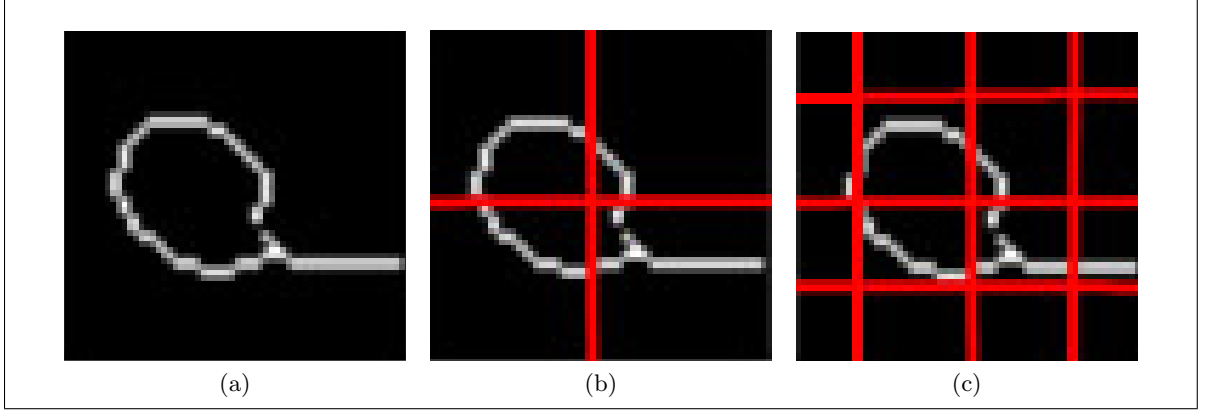


Figure 5: Results of applying PHOG onto saliency maps. (a) Level=0. (b) Level=1. (c) Level=2.

These classifiers were trained with information on appearance and shape, respectively, as we mentioned in section 2.3.

Second, the classifiers responses are kept in a vector with values belonging to $\{1, -1\}$. We summarize those values as a Rate of Votes (RoV) computed as the quotient between the sum of flaws and the sum of non flaws (1), where N is the number of views available for each flaw. This rate is used as a feature by a new classifier that then decides whether the sequence represents a true flaw. We calculated two rates: RoV_{shape} and $RoV_{appearance}$, from classification of shape and appearance, respectively. The combination of rates allows us to separate the classes correctly.

$$RoV = \frac{\sum_i^N Flaw Vote_i}{\sum_i^N Non Flaw Vote_i} \quad (1)$$

At this stage of the process we trained six kinds of classifiers: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Mahalanobis, Artificial Neural Networks (ANN), K-Near Neighbors (KNN), and (SVM) with Linear and Gaussian basis (Bishop et al., 2006). The best results were achieved by SVM and KNN.

3 Datasets and Methodology

We constructed a training dataset with 80 sequences as shown in Fig. 3, of those sequences 40 were true flaws and 40 non-flaws. The flaws were simulated utilizing the method proposed in Mery et al. (2005). We also constructed a test dataset composed of real flaw sequences previously unseen by the classifier in the training stages. The average number of views for each flaw sequence is 22 out of a possible 72 total views available from the analysis. Table 1 contains the specifications of each dataset.

The process starts off by extracting the features for shape and appearance, as mentioned in the section 2.3, individually from each element of the sequence. For each feature, we get a

Class	Number of Instances	
	Training	Test
Flaws	40	12
Nonflaws	40	37
Total	80	49

Table 1: Summary specification of dataset.

vector descriptor with U elements of length from every patch. Then, we put all the elements in the training set as a matrix of dimensionality $W \times U$, where W is the *total number of instances* $\times \sum$ *total number of views available in every trace* and U is the *length of the feature vector*. The same process is applied to the test dataset.

Features were separated into two groups, shape and appearance, as we described in section 2.3. We train two classifier SVMs for every group of features using ten fold cross validation (Bishop et al., 2006). We select the best performing classifier into the training process to use it in the following stages.

Next, each element of the sequence within the training set was classified individually as a flaw or non-flaw. We obtain the respective RoV_{shape} and $RoV_{appearance}$ according to equation (1) using the results of the classifiers. We trained six classifiers mentioned in the section 2.4: LDA, QDA, Mahalanobis, ANN, KNN, and SVM.

Finally, we tested our trained classifier with the dataset of real flaws, generating its corresponding RoV.

4 Experiments and Results

We evaluate our results by applying the standard two class analysis in pattern recognition based on estimating sensitivity (S_n) and specificity (S_p) as defined by equations (2) and (3). There are four possible outcomes of the classifier: TP, TN, FP and FN defined as True Positives, True Negatives, False Positives and False Negatives, respectively. Ideally $S_n = 1$ and $(1 - S_p) = 0$.

As we described in section 2.4, the first stage applies an individual assessment of the elements within the flaw sequence. The average performance of the classifiers in the training process was 89.7% for SVM_{shape} and 99% for $SVM_{appearance}$. We use F-Measure to select the best classifier in this stage, as defined in (4). This criterion allows us to characterize the performance in a single measure.

Classifier	S_n	$1 - S_p$	Correct Rate
LDA	0.8333	0.0000	0.9592
QDA	1.0000	0.4324	0.6735
Mahalanobis	1.0000	0.4324	0.6735
SVM linear	0.9167	0.0541	0.9388
SVM RBF $\sigma = 1$	0.9167	0.0541	0.9388
SVM RBF $\sigma = 2$	0.8333	0.0000	0.9592
SVM RBF $\sigma = 0.5$	0.9167	0.0811	0.9184
KNN	0.9167	0.0541	0.9388
ANN	0.9167	0.1351	0.9167

Table 2: Training of $SVM_{Appearance}$.

$$S_n = \frac{TP}{TP + FN} \quad (2)$$

$$1 - S_p = \frac{FP}{TN + FP} \quad (3)$$

$$F - Measure = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (4)$$

The final classifier was trained by applying RoV indicators, computed as mentioned in section 2.4 within the training set. The evaluation was conducted on a new testing set created with real flaws, previously unseen by the training process, both individually and in sequences. The best performance was obtained by the classifier (SVM) Linear, (SVM) with Gaussian kernels and $\sigma = 1$, and KNN. Overall, sensitivity is 92% and specificity is 95%. The complete evaluation is summarized in table 2.

The features relating to space were constructed with RoV indexes are shown in Fig. 6. It records the correlation between samples of testing and training although there is little difference between the training set and the testing set in the voting process on non-flaw samples.

5 Conclusions

In this research, we developed an approach for fault detection in aluminum castings based on object detection methods. Our approach maintains that by using a sliding-box it is possible to integrate information from multiple views and train a classifier with all available information, therefore ruling out a strict segmentation that has been used thus far. This approach allowed us to correctly detect faults without the need for finding correspondences as is classically seen in multiple view schemes.

The results obtained in the preprocessing stage of flaw detection based on saliency encourages the use of this method in non parametrical segmentation of flaws and reinforces

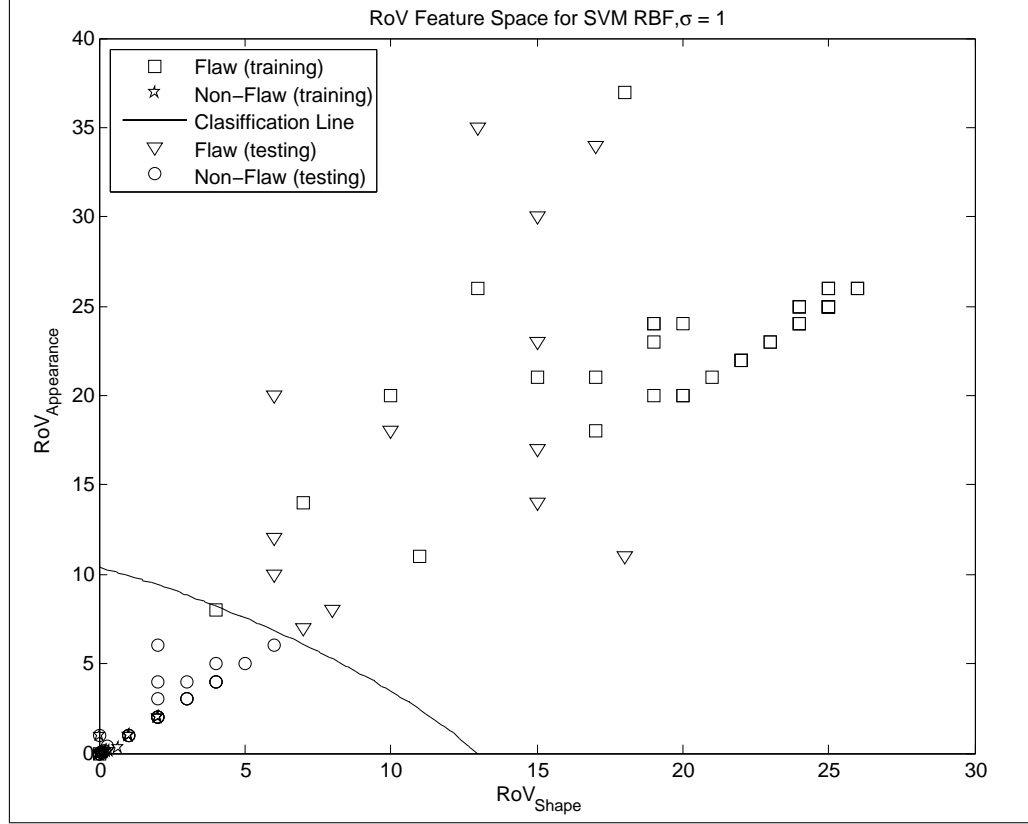


Figure 6: Feature Space for classifier SVM RBF with $\sigma = 1$.

our idea of emulating the human visual system of human operators. The detections were successful in the majority of the cases despite variations in the intensity of images, increasing the overall performance of the system.

Correlation between the classification of real samples and simulated ones suggests a novel form of learning which is useful in avoiding the issue of flaw detection when the flaws are uncommon and rare in actuality.

The methodology of machine learning along in addition to other tools gives our system the flexibility to deal with process requirements and therefore the capacity to adapt in the detection of various types of flaws.

We believe the results of our methodology are promising. However, we think that it is very possible to better take advantage of the information available from all views by utilizing models which are more complex. Based on this concept, we are working on a new model of information integration that allows us to define or discover a new structure of the object that we want to describe, that way we will be able to utilize our methodology in other detection problems.

References

- Achanta, R., Estrada, F., Wils, P., and Susstrunk, S. (2008). Salient region detection and segmentation. *Lecture Notes in Computer Science*, 5008:66.
- Bishop, C. et al. (2006). *Pattern recognition and machine learning*. Springer New York:.
- Boerner, H. and Strecker, H. (1988). Automated x-ray inspection of aluminum castings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(1):79–91.
- Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel. pages 401–408.
- Carrasco, M. and Mery, D. (2006). Automated visual inspection using trifocal analysis in an uncalibrated sequence of images. *Materials Evaluation*, 64(9):900–906.
- Dalai, N., Triggs, B., Rhone-Alps, I., and Montbonnot, F. (2005). Histograms of oriented gradients for human detection. 1.
- Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge Univ Pr.
- Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.
- Mery, D. (2003). Crossing line profile: A new approach to detecting defects in aluminium die castings. *Lecture Notes in Computer Science*, pages 725–732.
- Mery, D. and Filbert, D. (2002). Automated flaw detection in aluminum castings based on the tracking of potential defects in a radiosopic image sequence. *IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION*, 18(6).
- Mery, D., Filbert, D., and Jaeger, T. (2002a). Image processing for fault detection in aluminum castings. *Analytical Characterization of Aluminum and Its Alloys*.
- Mery, D., Hahn, D., and Hitschfeld, N. (2005). Simulation of defects in aluminium castings using cad models of flaws and real x-ray images. *Insight-Non-Destructive Testing and Condition Monitoring*, 47(10):618–624.

Mery, D., Jaeger, T., and Filbert, D. (2002b). A review of methods for automated recognition of casting defects. *INSIGHT*, 44(7):428–436.

Montabone, S. and Soto, A. (2009). Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing*.

Pizarro, L., Mery, D., Delpiano, R., and Carrasco, M. (2008). Robust automated multiple view inspection. *Pattern Analysis & Applications*, 11(1):21–32.

Purschke, M. (2002). IQI-sensitivity and applications of flat panel detectors and X-ray image intensifiers – a comparison. *Insight*, 44(10):628–630.

Viola, P. and Jones, M. (2004). Robust real-time object detection. *International Journal of Computer Vision*, 57(2):37–154.

APPENDIX B. HEAD MODELING

This appendix presents an article presented in The Chilean Workshop of Pattern Recognition 2012 that includes evaluation of the head classifiers using ensemble of features and ensemble of classifiers. This article handles the issue of integrate information from various points of views. We propose a multiple-view classification approach to bring a gap between advances in machine learning based object detection and multiple view geometry. The key idea is to classify an image sequence of corresponding parts of an object. This scheme allows us to solve problems related to correspondence throughout cameras, and to enhance the detection models with compounded features. We describe our approach applied in human head modeling by integration of visual information. The experiments demonstrate that our technique improves 2D state-of-art classifiers, using same training conditions. These results are promising and show that our approach can be use effectively to detect objects using multiple views.

Head Modeling Using Multiple-views

Christian Pieringer
cppierin@uc.cl

Domingo Mery
dmery@ing.puc.cl

Álvaro Soto
asoto@ing.puc.cl

Abstract

Object detection has attracted great interest of researchers in the computer vision community. Although machine learning approaches has been successful in this task, there are still significant challenges to solve in order to achieve data association, and including information from various points of views. We propose a multiple-view classification approach to bring a gap between advances in machine learning based object detection and multiple view geometry. The key idea is to classify an image sequence of corresponding parts of an object. This scheme allows us to solve problems related to correspondence throughout cameras, and to enhance the detection models with compounded features. This article describes our approach applied in human head modeling by integration of visual information. The experiments demonstrate that our technique improves 2D state-of-art classifiers, using same training conditions. These results are promising and show that our approach can be use effectively to detect objects using multiple views.

Keywords — Head detection, multiple views

1 Introduction

Object detection and recognition have been relevant research areas in computer vision along the last decade. The most relevant approaches based on machine learning categorize different kinds of objects using visual features extracted from image patches (Dalal and Triggs, 2005; Felzenszwalb et al., 2010; Viola and Jones, 2004). These researches focus on monocular scheme, and only few researchers have dedicated to exploit the use of multiple views to improve their performances. A few recent works focus in to demonstrate that 3D information, improve the detection. However, most of them include additional hardware, such as stereo cameras or depth cameras, due to they are focus on mobile robots.

In general, 3D recognition from 2D images is a complex task due to the infinite number of points of views and different illumination conditions (Poggio and Edelman, 1990). A simple recognition strategy consists in to performed by matching its invariant features with the features of a model. However, it may fail when objects have a large intra-class variation. In 2006, Rothganger et al. propose a novel representation for 3D objects is presented

based on local affine-invariant image descriptors and multi-view spatial constraints. The algorithm exploits the idea that smooth surfaces are always planar in the small. Thus, the matching and then the recognition is possible using photometric and geometric consistency constraints. A disadvantage is its poor performance on texture images. In Ferrari et al. (2006), a similar method based on the relationships among multiple model views enforces global geometric constraints in order to achieve 3D reconstruction from multiple views to recognizing single objects. A disadvantage is its poor performance on non textured images and uniform objects. In Voit and Stiefelhagen (2009), a tracking algorithm classifies the head pose base on the low resolution data in a multi-view camera system. An ellipsoid represents the head position and rotation, and a probabilistic framework joint the scores of the individual views. Finally, the tracking algorithm identify the real pose. There also is a different path for 3D object categorization, which use combination of information from multiple poses or points of view (Kushal et al., 2007; Savarese and Fei-Fei, 2007; Su et al., 2009; Thomas et al., 2006). However, they still are mono-focal classifiers. Although, 3D object classification and detection have had progress, especially linking features among views in a discriminative learning framework to create multiple view models of objects, there are still challenges to solve in order to improve data association.

We observe that mono-focal approaches for categorization suffer from: *i)* high efforts to improve classification models in only camera, and *ii)* discard available data from different visual sources. On the other hand, wide baseline stereo systems present an unsolved issue related to correspondence matching, where the same object has various poses or variations simultaneously. We propose an approach to categorize objects using simultaneous visual data, where the key idea is to use all the available visual information presents in a multi-view camera system, Fig. 1. The proposed approach offers several promising advantages in object categorization, including the following main contribution of this paper: improving classification performance using models on compounded visual features acquired simultaneously from the multi-view camera system. This framework let us to enrich the data used to train the models. Thus, we are able to include all the visual information in the same model.

This article presents our approach for people head modeling based on integration of visual information in a wide baseline stereo system. The results show our approach improve classification performance in average precision-recall, with a best performance when the algorithm use four cameras. These results are promising and demonstrate our approach can be used effectively to classify objects in multiple-views environments. The rest of the paper is organized as follows: Section 2 describes the proposed method. Section 3 provides implementation details, dataset details and main experiments of using our methodology in real images. Finally, Section 4 discuss concluding remarks and future avenues of research.

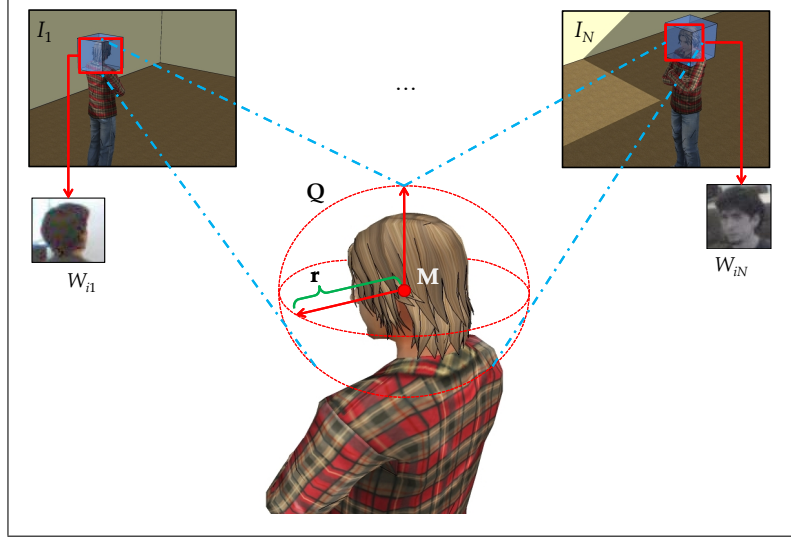


Figure 1: Diagram of head representation. We use N calibrated cameras C_1, \dots, C_N . In this example we assume the head is in positions $[X, Y, Z]$. The quadric \mathbf{Q} is projected from 3D space on images I_1, \dots, I_N to generate the windows W_1, \dots, W_N .

2 Overview of the Method

Similar to (Voit and Stiefelhagen, 2009), we assume that an object is represented by an ellipsoid. We use a quadric sphere located at coordinates $\mathbf{M} = [X, Y, Z]$ and radius r , which is totally defined as $\mathbf{Q} = (\mathbf{M}, r)$. We project this quadric onto the images in order to extract bounding-boxes where the object is located. After this process, we get a projected window W_j in the image j , as shown in Fig.2. A set of features represent each projection as inputs for a classifier. We decide over the joint data build up using all the projected windows W_j . More details about quadrics and conics representations can be found in Hartley and Zisserman (2003).

Our approach requires a fully calibrated multiple view system of N cameras C_1, \dots, C_N , with overlapped fields of view, to compute the geometric model which relates the 3D world homogeneous coordinates $\mathbf{M} = [X \ Y \ Z \ 1]^T$ to the 2D image coordinates $\mathbf{m}_j = [x_j \ y_j \ 1]^T$ in each image I_j . This model was obtained for $j = 1, \dots, N$ cameras using the transformation $\lambda \mathbf{m}_j = \mathbf{P}_j \mathbf{M}$, where λ is a scale factor, and \mathbf{P}_j is the 3×4 calibration matrix of camera C_j (Hartley and Zisserman, 2003).

2.1 Feature Extraction

We rescale each projection W_j to 64×64 pixels to cope with different sizes, and we extract a set of features in pyramidal decomposition for each window (Bosch et al., 2007; Lazebnik et al., 2006). This allows us to represent global and local information from each object instance. Each level $l \in L = \{0, \dots, n\}$ in the pyramid has 4^l cells or patches, and for each

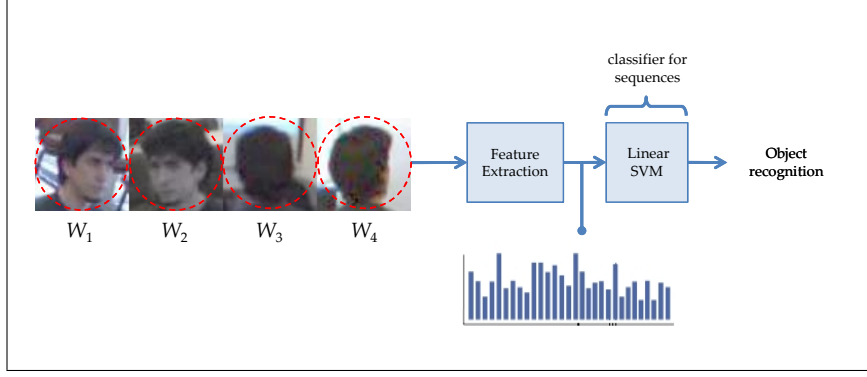


Figure 2: Block diagram of the proposed method. Our approach includes two main steps: feature extraction, sequence classification. The algorithm begins with a input sequence composed by W_1 to W_N . This sequence represents the quadrics seen from each camera view. We draw the projection Q as dashed red circles to show graphically how to select the maximum parallelepiped subscribed to Q . Each element W_j was cropped, and then rescaled to 64×64 pixels to cope with projection at different size. We extract LBP features for each projection W_j , and finally apply the model in order to classify the sequences.

cell we compute a descriptor with K bins. The descriptor of the entire image patch W_j has $N_f = K \sum_{l=0}^L 4^l$ bins. As recommended in Bosch et al. (2007), we use $L = 3$.

We use Local Binary Patterns (LBP) proposed in Ojala et al. (2000) as a measure of texture that uses local appearance descriptors. It is computed comparing a center pixel with its neighbors and this comparison is represented as decimal number. The final LBP descriptor contains $K = 59$ bins. This feature outperform HOG in clutter backgrounds and different textures (Wang et al., 2009), such as different poses of the head within the patch sequence, as shown Fig. 2.

2.2 Classifier for Sequences

Once we extracted features on each element W_j , we apply two independent and exclusive scheme each other to classify the projections sequences: gathering features and ensemble of classifiers. Along the experiments, we evaluate both approaches in order to present pros and cons of them. In both cases, we use support vector machines (SVM) with linear kernels as classifiers Cortes and Vapnik (1995). SVM with linear kernels improves the classification accuracy and speed andover SVM with no-linear kernels in image categorization problems (Yang et al., 2009).

As we mentioned previously, both are independent and exclusive each other. A bootstrap strategy, similar as Felzenszwalb et al. (2009), allows to avoid memory overloads and overfit along the training. We let the bootstrap algorithm picks a ratio of misclassified samples and releases a ratio of well classified for the next training round. The next two sections describe how we trained both classification schemes for projections sequences: gathering features and ensemble of classifiers.

Table 1: Details of train dataset used for training individual models

class	number of examples
frontal head	916
rear head	916
background	9.583
total	11.415

2.2.1 Gathering Features

In this scheme, we train a single linear SVM classifier with features concatenated as single descriptor, *i.e.*, each W_j have associated a feature vector with N_f bins and the camera system has N_{cam} cameras. Then, the sequence descriptor $N_s = N_{cam} \times N_f$ elements. The key idea is to build up an enriched feature vector, which represent the head structure in a global perspective. After the training, we get a model SVM_{gather} that will be used to classify whole the sequence.

2.2.2 Ensemble of Classifiers

The ensemble of classifier is composed by two layers. In the first layer, three individuals linear SVM models, β_1 , β_2 and β_3 , learn to discriminate among three classes respectively: frontal heads, rear heads, and background. All of them learn in one-vs-all fashion. In the second layer, a new linear SVM model use the scores from the previous layer to discriminate the whole sequence. The three scores, f_{β_1} , f_{β_2} , f_{β_3} build up the feature vector at the second layer, which has $N_s = N_{scores} \times N_{cam} = 3 \times N_{cam}$ elements. This final classifier $SVM_{ensemble}$ is able to merge the information coming from the camera system.

3 Experiments and Results

In this section, we describe implementation details and results to applying our approach in the classification task.

3.1 Dataset Details

We build our own multi-view head dataset for training and testing using our camera system, due to the lack of multi-view datasets for people head or people torso. This camera system consists of four synchronized cameras. All images were acquired at 640×480 pixels and 15fps. We manage two train datasets: one for training individual models used for evaluate each projection W_j , and one for training the ensemble.

In both train datasets, we use a set of 10 people placed within a room, spinning over their Z axis from 0° to 360° . Negative samples include objects such as clothes, computers, walls. We also combined individual samples randomly in order to build artificial negative

Table 2: Details of train dataset used for training the ensemble classifier

class	number of examples
frontal head	233
background	3.108
total	3.441

sequences and enrich the sequence dataset. The test dataset was formed by 300 frames fully labeled from two multi-view video sequences in a classroom or auditorium environment, where people were sat and following a speaker. Sequences are manually labeled in the four cameras. People in this dataset are different to people who appears in the train dataset.

3.2 Experiments

We evaluate our approach using the both classification schemes. Experiments measure the ability to improve the discriminative power, and centering ability.

3.2.1 Enriched features

During the training process, we evaluated the influence of adding information coming from more visual sources. We apply the analysis in terms of classification performance. We started training a classifier only using data from one camera and test this model using the test dataset. We repeated this process along as we add more visual sources. We observe an increase in performance from 40% to 70% of average precision-recall, with maximum considering the four views, as shown Fig 3. As we stated, using more cameras we enhance the features with the complementary information available in the other cameras.

3.3 Centering

Once we trained both schemes of classifiers, we pick centered and non-centered projections. In Fig. 4, we show three sets of four candidate sequences. The first and second sets, Fig.4a and Fig. 4b belong to head class, and the third set , Fig. 4c belongs to the background class. The classifier for sequences have the main task to discard sequences belong to the background, but we note this model intrinsically also do the task to align the sequence as it learnt in the training process. The higher scores were always given to the best alignments as shown the best scores. All the scores in the third set c) are strongly negative, and therefore all assigned to the background class.

4 Conclusions

We proposed a head classifier based on wide-baseline stereo camera system. Our approach showed a main contributions of improving classification performance using models on com-

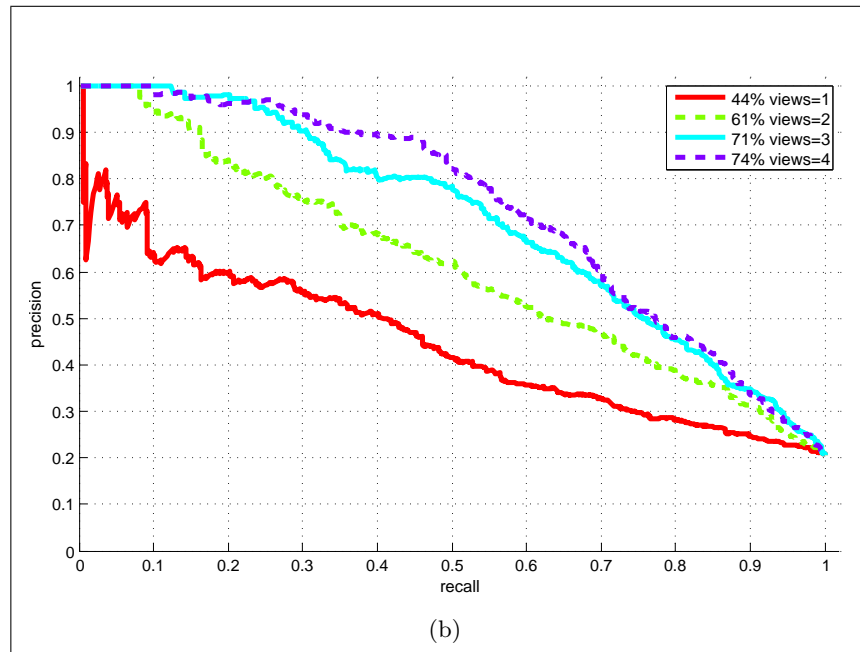
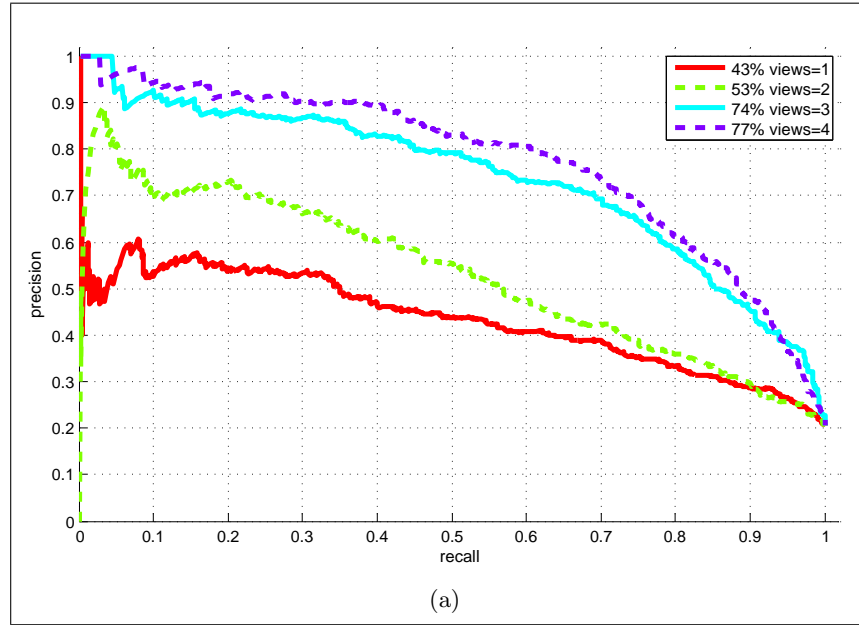


Figure 3: Both curves show the performance evolution by adding information from various visual sources. Fig. (a) and (b) shows the gather and ensemble strategies, respectively.



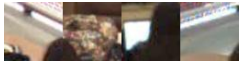

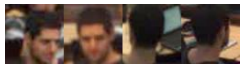
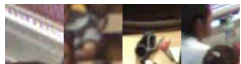
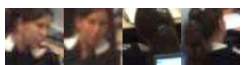
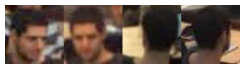

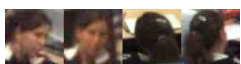

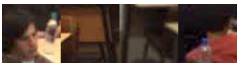
Image Example	Score	Image Example	Score	Image Example	Score
	-3.591		0.74		-2.20
	-2.58		1.24		-3.28
	-1.84		1.66		-4.54
	1.17		-1.92		-3.83
(a)		(b)		(c)	

Figure 4: Image sequences resulting of pass sliding-box among a set of neighbors, and them scores. Box scores are high when the box belongs to head class, and its projections reach the better alignment, as shown in (a) and (b). In (c) we observe background examples and their score, all negatives.

pounded visual features acquired simultaneously from the multi-view system. Both classification schemes show similar behaviors performances. Although we did not address occlusion issues, our experiments showed promising results using information from various points of view in the same scene. The integration of information through our approach is able to codify an structure inherent to the image sequence and therefore an object structure, in this case, head structure. The ensemble also works as a case parts-based approaches, which codify this mid-level structures. One disadvantage is the calibration process, which makes to our approach somewhat rigid to the scene structure. We believe our results are promising, and our approach can be adapted for a another challenging multi-view scenario. For future work, we plan to address occlusion issues and cases where it missing projections within the sequence. We believe possible extract information coded in the sequence which reveal if certain images do not belong to the same window detection. We would like also address the head pose estimation problem using the same framework.

Acknowledgment

This work was supported by the ACT-32 Project focused to improve public transport systems, and Fondecyt N. 1100830 project.

References

- Bosch, A., Zisserman, A., and Muñoz, X. (2007). Image classification using random forests and ferns. *IEEE International Conference on Computer Vision (ICCV)*.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. *Proc. CVPR*, 1:886–893.
- Felzenszwalb, P. F., Girshick, R. B., and McAllester, D. (2010). Cascade object detection with deformable part models. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2009). Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints).
- Ferrari, V., Tuytelaars, T., and Van Gool, L. (2006). Simultaneous object recognition and segmentation from single or multiple model views. *International Journal of Computer Vision*, 67(2):159–188.
- Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*. Cambridge University Press.
- Kushal, A., Schmid, C., and Ponce, J. (2007). Flexible object models for category-level 3d object recognition. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*.
- Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proc. CVPR*, 2(2169-2178):1.
- Ojala, T., Pietikainen, M., and Maenpaa, T. (2000). Gray scale and rotation invariant texture classification with local binary patterns. *Lecture Notes in Computer Science*, 1842:404–420.
- Poggio, T. and Edelman, S. (1990). A network that learns to recognize 3d objects. *Nature*, 343(6255):263–266.
- Rothganger, F., Lazebnik, S., Schmid, C., and Ponce, J. (2006). 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, 66(3):231–259.
- Savarese, S. and Fei-Fei, L. (2007). 3d generic object categorization, localization and pose estimation. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*,

pages 1–8.

Su, H., Sun, M., Fei-Fei, L., and Savarese, S. (2009). Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In *International Conference on Computer Vision*.

Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., Schiele, B., and Van Gool, L. (2006). Towards multi-view object class detection. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1589–1596.

Viola, P. and Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154.

Voit, M. and Stiefelhagen, R. (2009). A system for probabilistic joint 3D head tracking and pose estimation in low-resolution, multi-view environments. *Computer Vision Systems*, pages 415–424.

Wang, X., Han, T. X., and Yan, S. (2009). An hog-lbp human detector with partial occlusion handling. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 32 –39.

Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794 –1801.

APPENDIX C. HEAD DETECTION USING SLIDING-BOXES IN MULTIPLE VIEWS

This appendix presents a draft article about head detection using our *sliding-box* approach. This article summarizes the results of this thesis. Sliding-Window Detectors have attracted great interest among researchers in the computer vision community because of the advantage they offer of avoiding segmentation problems during detection. We propose a generalization of the sliding-windows approach to 3D cases, which we call sliding-box. This approach works on calibrated multiple view configurations where we have various viewpoints of the same scene. This calibrated system allows us to run a sliding-box in world coordinates and project this box in its corresponding position on each image. We can also search in the correct scale and location, improving the processing limitations of the sliding-box. Furthermore, it allows us to create appearance models from various viewpoints to improve detection. We apply our approach to head detection as the head is useful when the rest of the body is occluded. Experiments show that our framework improves detection performance by 10% of average precision-recall as compared to the optimal view of state-of-the-art 2D methods in our datasets. These results suggest that our approach can be used effectively to detect objects in multiple view systems, improving detection performances achieved by 2D detectors in isolation.

Heads Detection Using Sliding-boxes in Multiple Views

Christian Pieringer
cppierin@uc.cl

Domingo Mery
dmery@ing.puc.cl

Álvaro Soto
asoto@ing.puc.cl

Abstract

In the context of object detection, the sliding-window method has become the favorite approach to apply object classifiers to 2D images. As a major advantage, this method avoids the need for a segmentation step that usually leads to poor results. As a major limitation, it performs a brute-force search for object detections at image positions and different scales, usually leading to a high computational load. In this work, we argue and demonstrate that a suitable use of geometric constraints can play a key role to improve this search scheme, leading to higher levels of recognition accuracy and a substantial reduction in the number of candidate window locations. Furthermore, it is possible to avoid false positive detections arising in image areas corresponding to background clutter. In particular, we propose a generalization of the sliding-window method to 3D cases, where we slide a 3D box in world coordinates instead of a 2D window in the image plane. Consequently, we call this a sliding-box approach. This approach operates on a calibrated multiple view configuration, where we acquire several viewpoints of the same scene. This calibrated system allows us to run a sliding-box in world coordinates and project this box into its corresponding position in each 2D image. As a major advantage, general knowledge about the target object allow us to slide the box at a suitable scale and environment positions. Furthermore, it is possible to train object classifiers using integrated multiple view appearance features. We apply the proposed approach to the task of head detection in indoor environments. Our experiments show that, in terms of average precision-recall, the proposed framework improves detection performance by 10% with respect to approaches based on a single view detection.

Keywords — Sliding-window, head detection, LBP, multiple views

1 Introduction

The success of appearance methods based on visual descriptors and off-the-shelf machine learning techniques (Dalal and Triggs, 2005; Dean et al., 2013; Felzenszwalb et al., 2010b; Girshick et al., 2013; Viola et al., 2005; Yang et al., 2009) is one of the main reasons of the new enthusiasm for visual recognition technologies. These methods have shown robustness against visual complexities such as variations in illumination, scale, affine distortions, and

mild intraclass and pose variations. Although there have been notable progress, overall performance is still poor (Dollar et al., 2011).

A common denominator of these techniques is that they rely mainly on statistical learning methods that exploit image-intensity information to capture object appearance features. Their goal is to reveal visual spaces where visual similarities carry enough information to obtain robust visual recognition. As a relevant limitation, appearance-based approaches do not consider geometric information that can provide key constraints to reduce the search space of possible objects locations and scales. Some notable exceptions with promising results are Espinace et al. (2013); Salas and Tomasi (2011); Spinello and Arras (2011); however, these approaches require range sensors to capture geometric information.

In general, we observe that single view approaches to object detection mainly *i)* use a sliding-window at various scales to compensate scale changes of the object target class in images, resulting in false positives due to hallucinations at several scales; and *ii)* do not take into account useful 3D information such as real sizes of people or objects, and the positions in which they are likely to be found in the scene. Fortunately, geometric techniques exist for establishing relationships across views in a camera system, providing useful ways of combining and integrating this information (Hartley and Zisserman, 2003; Szeliski, 2010). This facilitates the use of geometric information. Furthermore, it provides a higher level of robustness to occlusion problems in single views.

In this work, we propose a method for detecting people in a calibrated multiple view system in which a 3D sliding-box is tailored to the physical size of the target object class. This box \mathbf{B} is applied according to the three directions ($X; Y; Z$) of the relevant world frame where people are likely to appear. Our approach is designed to inspect the corresponding portion of the images where the volume contains projections, on the basis of size as shown in Fig. 1. A calibration model provides a geometric relationship between world coordinates and images, and relationships between cameras in multiple view configurations. This geometric structure allows us to filter out false detections present in one view when detections are not consistent with the remaining views. In addition, knowledge about the target object and the geometry of the scene allow us to reduce the search space, sliding the box only at a suitable scale and likely locations.

The proposed approach offers several promising advantages for object detection, including the following main contributions:

1. By exploiting geometric constraints, it can use information about object size and scene configuration to significantly reduce the search space of a traditional sliding-window approach, leading to higher levels of recognition accuracy and a reduction in the rate of false positive detections.
2. By projecting the 3D box to the corresponding image planes, it can check detection

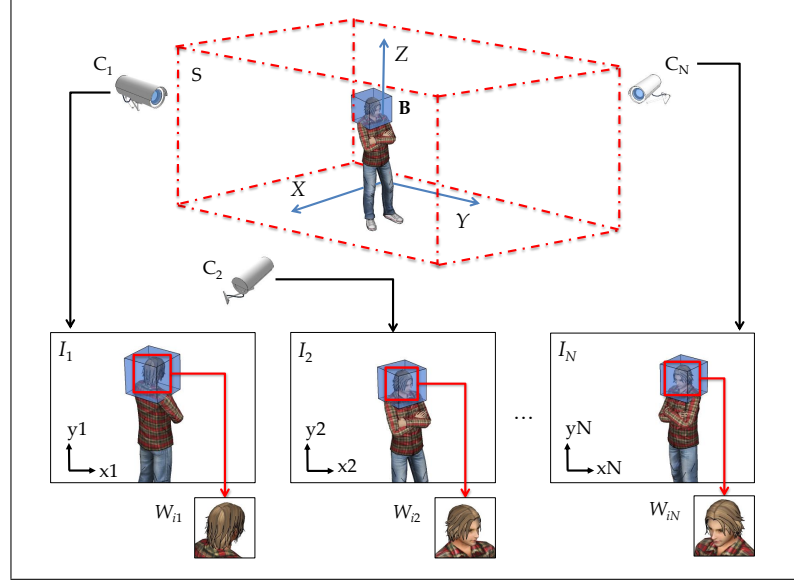


Figure 1: Diagram of the proposed approach for detecting people using a multiple view system. This approach requires N calibrated cameras C_1, \dots, C_N . In this example, the sliding-box \mathbf{B} is placed at various positions $(X; Y; Z)$, scanning the space S of the scene in which heads could be located. The sliding-box B is projected from 3D space onto images I_1, \dots, I_N , retrieving N detection windows W_{i1}, \dots, W_{iN} . As key advantages, geometric information allows us to directly scan the image at suitable locations and scales for human heads.

consistency in the 2D views, leading to the elimination of spurious detections.

3. By integrating in a single object descriptor appearance features from multiple image views, it can provide a classification scheme more robust to pose and occlusion problems.
4. By applying the proposed approach to the task of people-counting using head detection, it is possible to verify the relevance of the previous ideas.

This paper is organized as follows. Section 2 presents a review of previous work related to people detection in single and multiple views. Section 3 provides a detailed description of our approach. Section 4 describes the datasets generated to test our approach. Section 5 presents our experimental evaluation. Finally, Section 6 presents conclusions related to our main results and discusses future research avenues.

2 Related Work

Category based object detection is a very active topic in the computer vision literature with an extensive list of previous approaches (Dalal and Triggs, 2005; Dean et al., 2013; Dollár et al., 2010; Felzenszwalb et al., 2010a; Girshick et al., 2013; Papageorgiou and Poggio, 2000; Pedersoli et al., 2014; Tang et al., 2013; Viola and Jones, 2001). In particular, techniques

for people detection have been highly popular, as there are many potential applications that could benefit from this technology. We defer the reader to Dollar et al. (2011) for a recent survey about this topic. Head detection is a special case in people detection. Important conclusions from works on people detection suggest that detecting the head helps to avoid occlusions in crowded environments, as the head is the least frequently occluded part of the body under such conditions. Also, heads are less deformable than the rest of the body (Ali and Dailey, 2012; Dalal and Triggs, 2005; Eshel and Moses, 2010). Most previous works on head detection incorporate the detectors into more complex detections systems. In particular, as part-based detector that helps to find the complete body or other body parts, thus improving detection (Ali and Dailey, 2012; Chang et al., 2013; Hayashi et al., 2013; Nghiem et al., 2012; Xie et al., 2012; Zeng and Ma, 2010).

While most previous works on people and head detection have been based on a single camera case, there have been also works that address the multiple view case. In general, previous works demonstrate the ability of multiple views to improve detection performance over single view cases (Ali and Dailey, 2009; Delannay et al., 2009; Pane et al., 2013). Most current works on multiple view detections use a calibrated and overlapped set of cameras, where background subtraction techniques are used to detect targets in each single view. Afterwards, a ground-plane homography is used to merge these single view detections. Finally, a multiple view tracking algorithm discards false hypotheses using visual appearance information (Ali and Dailey, 2009; Delannay et al., 2009; Eshel and Moses, 2010; Kim and Davis, 2006; Liem and Gavrila, 2013).

There are significant works on representing 3D objects using 2D images, based on the psycho-physical premise that a 3D structure can explain all of the changes in appearance arising from changes of viewpoint or aspect of the object, as is done by the human brain (Mundy, 2006; Stone, 1999; Tarr and Kriegman, 2001). Most of these approaches use an aspect-graph representation to establish relationships among the different topological appearances of a target object. In this graph, a node represents object views that are adjacent on the unit sphere of viewing, while an edge arises from the transition in the graph that relates the different faces or views of the projected object. Although, these methods are trained using data from various viewpoints, they generally still work using a single view scheme during testing (Cyr and Kimia, 2004; Koenderink and van Doorn, 1979; Savarese and Fei-Fei, 2007). Furthermore, these models present several practical disadvantages: *i*) the size of the aspect-graph grows rapidly with the topological transitions required for object recognition, meaning that the aspect-graph becomes application-specific (Mundy, 2006); *ii*) the scale of the viewing distance required to determine the relevant transitions in accordance with the topology of the object is not known in advance, thus small scale transition that are topologically significant may not be relevant for the object recognition (Mundy, 2006); and, *iii*) Complexity of aspect generation, storage, and search requirements are impractical

for objects of modest complexity (Cyr and Kimia, 2001).

From the previous review, we note that head detection is a useful cue during people detection in indoor environments, therefore we focus our work on this application. Our framework also explores the properties of multiple view configurations that improve detection. However, instead of directly using a matching technique to relate detections, we use the simultaneous projections of the target object into the views. We also note that aspect-based representation fits well with multiple view environments where people’s aspects are acquired from various viewpoints simultaneously. Furthermore, using this representation in a multiple view framework might include enriched appearance information and 3D cues to help locate people in the scene. We include this idea of aspect-graph in terms of collection of aspects that describe an object from various viewpoints.

3 Proposed Approach

We propose a 3D extension of the single view sliding-window approach. Specifically, we propose a generalization of the sliding-window method to 3D cases, where we slide a 3D box in world coordinates instead of a 2D window in the image plane. Consequently, we call this a sliding-box approach. At each 3D position in world coordinates, an sliding-box defines a 2D projection onto the image plane of each camera covering the scene. This set of projections forms a collection of aspects (Cyr and Kimia, 2004), as shown in Fig. 2. The term “aspect” refers to prototypical views or templates of an object that are similar to each other. In this sense, if an object occupies the space covered by the sliding-box, we expect to observe a high level of consistency among the appearance descriptors of the set of 2D projections. This scheme allows us to simultaneously consider the information in the multiple view camera system in order to filter-out false positive detections on single views. Furthermore, previous knowledge about the size and likely location of the object of interest, in our case heads, provide further constraints to slide the box in suitable locations.

We take advantage of the previous properties of a sliding-box approach by proposing a new method based on five main steps: spatial focus-of-attention, multiple view projection, feature extraction, multiple view classification, and non-maximal suppression, as shown in Fig. 2. Next, we describe the details behind each of these steps.

3.1 Spatial Focus-of-attention

A major problem in applying the proposed sliding-box method is the great computational complexity of projecting each box onto the different views. Unfortunately, previous strategies to reduce the search space, such as coarse-to-fine search schemes (Pedersoli et al., 2010), are not very effective in the 3D case of a sliding box. In this work, we propose to reduce the search space by applying a pre-processing step that consists of using a salient detector and

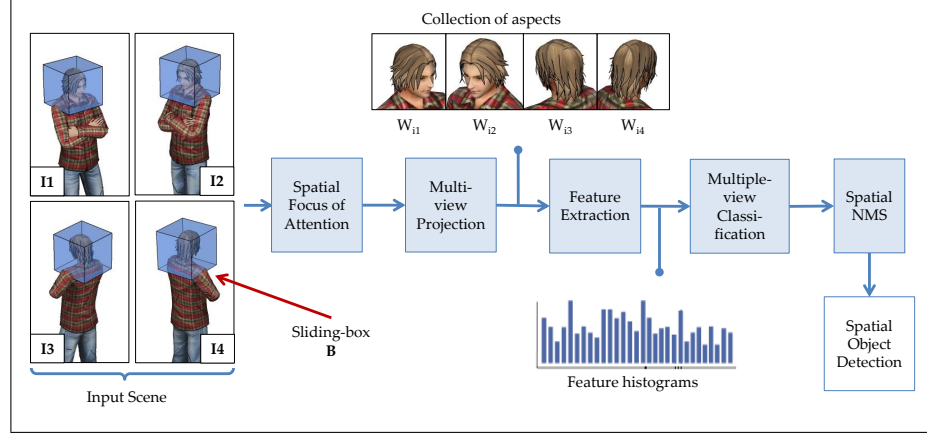


Figure 2: Block diagram of our proposed method for the case of $N = 4$ cameras. As shown, our method consists of five main steps. First, we apply a spatial focus-of-attention mechanism to pre-filter false detection and to reduce the search space of candidate sliding-boxes. Then, we project each sliding-box \mathbf{B}_i onto each image I_j forming a collection of aspects W_{ij} . Next, we extract features from each projection W_{ij} . Afterwards, we apply a multiple view classifier to the collection of aspects. Finally, an NMS procedure allows us to eliminate multiple detections.

multiple view geometry to generate a focus-of-attention in 3D space. First, we run a standard 2D sliding-window detector as a specialized region detector to find head hypotheses in the 2D image views, as shown in Fig 3a. Let \mathbf{h}_{ij} be the image coordinates for the centroid of head hypothesis i in image j . We match hypothesis \mathbf{h}_{ij} with the set of hypotheses \mathbf{h}_{rs} closer to the epipolar line $\mathbf{l}_s = \mathbf{F}_{js} \cdot \mathbf{h}_{ij}$ in image s , where \mathbf{F}_{js} is the fundamental matrix estimated using camera matrices. Then, the spatial location $\hat{\mathbf{M}}_i$ for each pair of matched hypotheses \mathbf{h}_{ij} and \mathbf{h}_{rs} is triangulated by a least square minimization using the corresponding image coordinates in 3D space (Hartley and Zisserman, 2003). Finally, we select as candidate locations for image heads the set of neighboring points around the estimated position $\hat{\mathbf{M}}_i$ with the likely location of the true heads, as shown in Fig. 3b. These sets of hypotheses lie close to the ground-truth heads, as show in Fig. 3c.

3.2 Multi-view Projection

A sliding-box \mathbf{B} has three properties: *i)* \mathbf{B} is centered at 3D point $\mathbf{M} = [X \ Y \ Z \ 1]^T$ using homogeneous coordinates; *ii)* \mathbf{B} belongs to the 3D space of interest S where the target object class can be located; and, *iii)* \mathbf{B} contains a volume that is tailored to the real size and shape of the object classes to be detected. In our case, we are detecting heads. For practical purposes, the i -th sliding-box \mathbf{B}_i defines a sphere circumscribed to its occupied space, centered in \mathbf{M}_i and with radius $r = 15[cm]$. This geometric representation fits well with the oval shape of heads. The sphere is algebraically expressed as a quadric surface defined in the world frame, which projects a conic section \mathbf{C} on an image I_j (Hartley and Zisserman, 2003), as shown in Fig. 4.



Figure 3: Focus-of-attention procedure. (a) Shows the triangulation between head hypotheses in two images I_1 and I_2 from different viewpoints. Hypotheses $\{h_{11}, h_{21}, h_{31}\}$ in image I_1 generate epipolar lines l_1, l_2 and l_3 in image I_2 . The pairs of head hypotheses $\{h_{11}, h_{12}\}$ and $\{h_{21}, h_{22}\}$ share the same 3D position \hat{M}_1 and \hat{M}_2 , respectively. We estimate these spatial positions by triangulation using least square minimization along the ray $h_{ij}C_i$. As there are no head hypotheses near to epipolar line l_3 , the h_{31} does not generate a potential head position. (b) Blue dots show the potential head positions detected by our focus-of-attention procedure. This process provides hypotheses with more likely location of heads within the region of interest S and helps us to drastically filter spatial detections. (c) White circles show examples of ground-truth heads within the subspace S (best viewed in color).

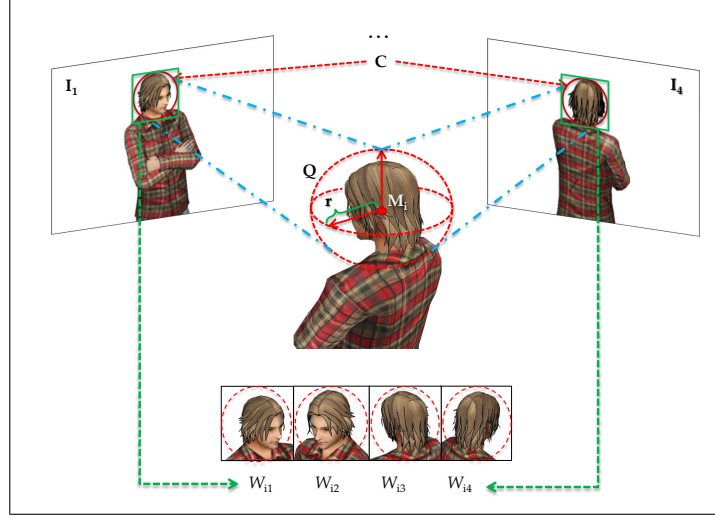


Figure 4: Projection diagram of a sphere quadric \mathbf{Q} defined on \mathbf{M}_i with radius r . In this example, for $N = 4$, \mathbf{Q} is projected onto the images I_1, \dots, I_4 as a conic \mathbf{C} . The projections W_{ij} are defined as the maximum quadrilateral subscribed over \mathbf{C} (dashed red circles). The elements W_{ij} represent the projection of \mathbf{B}_i onto camera j . All of these elements define a collection of aspects which represents box \mathbf{B}_i seen from each camera. Each element W_{ij} is cropped and then rescaled to 64×64 pixels before feature extraction to cope with projection at different sizes (best viewed in color).

In this step, we compute the sliding-box projections onto the 2D images in order to generate correlated image sections. Our approach requires a fully calibrated multiple view system of N cameras, C_1, \dots, C_N , to compute a geometric model that relates world coordinates \mathbf{M}_i to 2D image coordinates $\mathbf{m}_{ij} = [x_{ij} \ y_{ij} \ 1]^T$ in each image I_j . This model is obtained for $j = 1, \dots, N$ cameras using the transformation $\lambda \mathbf{m}_{ij} = \mathbf{P}_j \mathbf{M}_i$, where λ is a scale factor and \mathbf{P}_j is the 3×4 calibration matrix of camera C_j (Szeliski, 2010). We let the sliding-box runs within the space of interest S , defined in 3D as a rectangular parallelepiped where coordinates X , Y , and Z are constrained by a lower and a higher boundary a and b , as shown in Fig. 5. This top-down information allows us to limit the action of our sliding-box to areas where we expect to find heads in the scene, e.g., we do not expect to find heads on the ceiling or lying on the floor, as shown in Fig. 5.

3.3 Feature Extraction

We represent each window W_{ij} as a bounding-box in I_j defined as the maximum subscribed quadrilateral in the sphere's projection of box \mathbf{B}_i corresponding to the conic \mathbf{C} . We extract a set of features using a pyramidal decomposition for each rescaled version of window W_{ij} (Bosch et al., 2007; Lazebnik et al., 2006), where each W_{ij} is represented by a feature vector \mathbf{x}_{ij} . This strategy allows us to extract global and local spatial information from each head instance. Each level $l \in L = \{0, \dots, n\}$ in the pyramid has 4^l cells or patches, and for each cell we compute a descriptor with K bins. The descriptor of the entire image patch \mathbf{x}_{ij} has

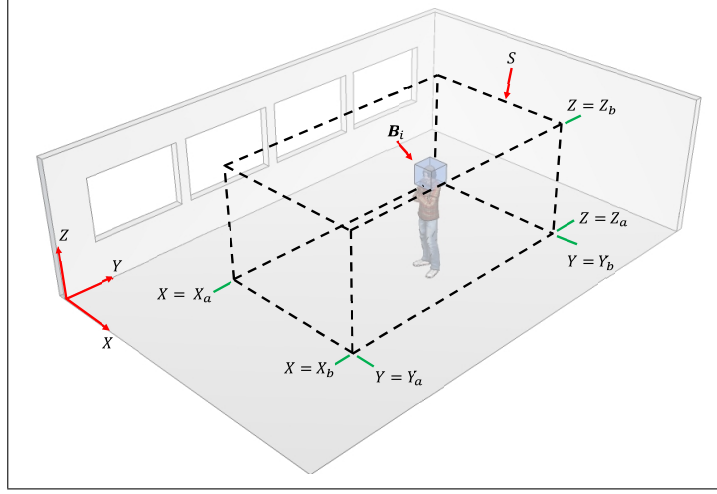


Figure 5: Diagram of the space of interest S inside of a room and defined as a parallelepiped with set of boundaries $[X_a, X_b]; [Y_a, Y_b]; [Z_a, Z_b]$. We use this contextual information to limit the action of our sliding-box \mathbf{B}_i within the space S . This allows to us to search for people’s heads in areas in which they are likely to appear according to context.

$N_f = K \sum_{l=0}^L 4^l$ bins. In this work we build the descriptor using Local Binary Patterns (LBP) (Ojala et al., 2000) as low level features.

3.4 Multiple View Classifier

We apply two different strategies to classify the collection of aspects provided after the multiple view projections: ensemble of features and ensemble of classifiers. In both cases, we use Support Vector Machines classifiers (SVM) (Cortes and Vapnik, 1995). The next two sub-sections describe the two multiple view classification approaches.

3.4.1 Ensemble of Features (EF)

This classification procedure includes the main steps summarized in Fig. 6. First, we concatenate the features of each aspect W_{ij} into a single descriptor $\mathbf{x}_i = [\mathbf{x}_{i1} \dots \mathbf{x}_{iN}]$ with dimension N_f bins. This representation allow us to integrate and to simultaneously evaluate the collection of aspects defined by the projections of the sliding-box. The key idea is to build up an enriched feature vector that represents the head structure, as shown in Fig. 4. Then, we train a single SVM classifier, which scores an instance \mathbf{x}_i with a function of the form $f_{\beta}(\mathbf{x}_i) = \beta \cdot \Phi(\mathbf{x}_i)$, where β is the vector of model parameters and Φ is a kernel function to transform the example \mathbf{x}_i to a high dimensional space to separate the target classes.

3.4.2 Ensemble of Classifiers (EC)

As an alternative to the previous classification scheme, Figure 7 shows the main steps of our classification procedure using a two-layers classifier ensemble. First, each aspect W_{ij}

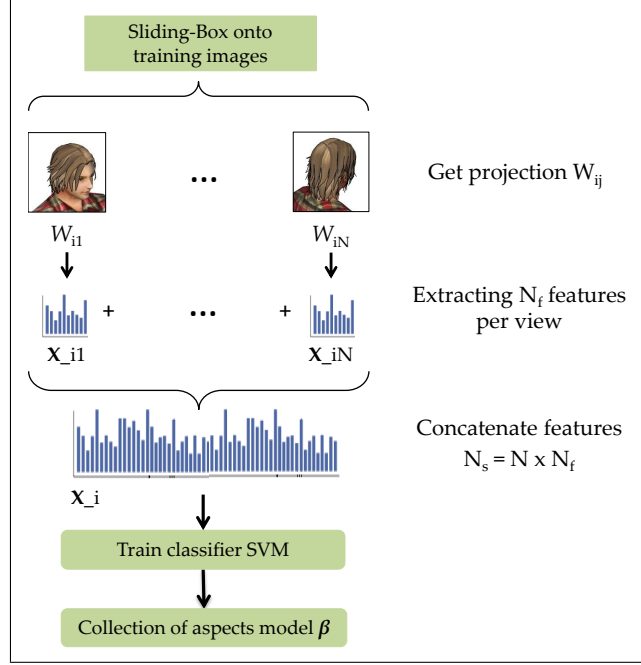


Figure 6: Training process diagram of features ensemble scenario. Once we extract features from each element W_{ij} , we concatenate all N_f features in a single descriptor with N_s bins. We use a single SVM classifier to learn a model β using examples of head sequences.

is described by feature vector \mathbf{x}_{ij} . Then, we use each 2D view to train linear multi-class SVM model to discriminate among k subclasses and assigns a confidence score $f_{\beta}^k(\mathbf{x}_{ij})$ to each class. This layer of classifiers transform the set of feature vectors \mathbf{x}_{ij} into a set of mid-level features such that $\mathbf{s}_{ij} = [f_{\beta}^1(\mathbf{x}_{ij}), \dots, f_{\beta}^k(\mathbf{x}_{ij})]$. The subclasses represent head aspects and background. Finally, in the second layer we train a new SVM model that uses the concatenation of scores \mathbf{s}_{ij} as single descriptor $\mathbf{s}_i = [f_{\beta}^1(\mathbf{x}_{i1}) \dots f_{\beta}^k(\mathbf{x}_{iN})]$, where N is the number of views used in the detection process. This new descriptor identifies the entire collection of aspects. We address the SVM multi-class problem using two methods: one-against-all (OVA-EC) and one-against-one (OVO-EC). During training we follow a standard practice in order to improve the precision of each classifier by adding a bootstrapping step to mine hard negative examples. Specifically, starting from an initial classifier, we conduct an iterative process where at each iteration we re-train the current classifier using an improved training set that includes previous false detections of hard examples. Finally, this procedure generates the model used to classify the collection of aspects during the detection process.

3.4.3 Best Collection of Aspects

Classifiers described in Sections 3.4.1 and 3.4.2 discriminate over aligned collection of aspects, *i.e.*, the aspects follow a specific sequence of appearance models according to the pose of the target object in relation to each camera. However, during detection we do not

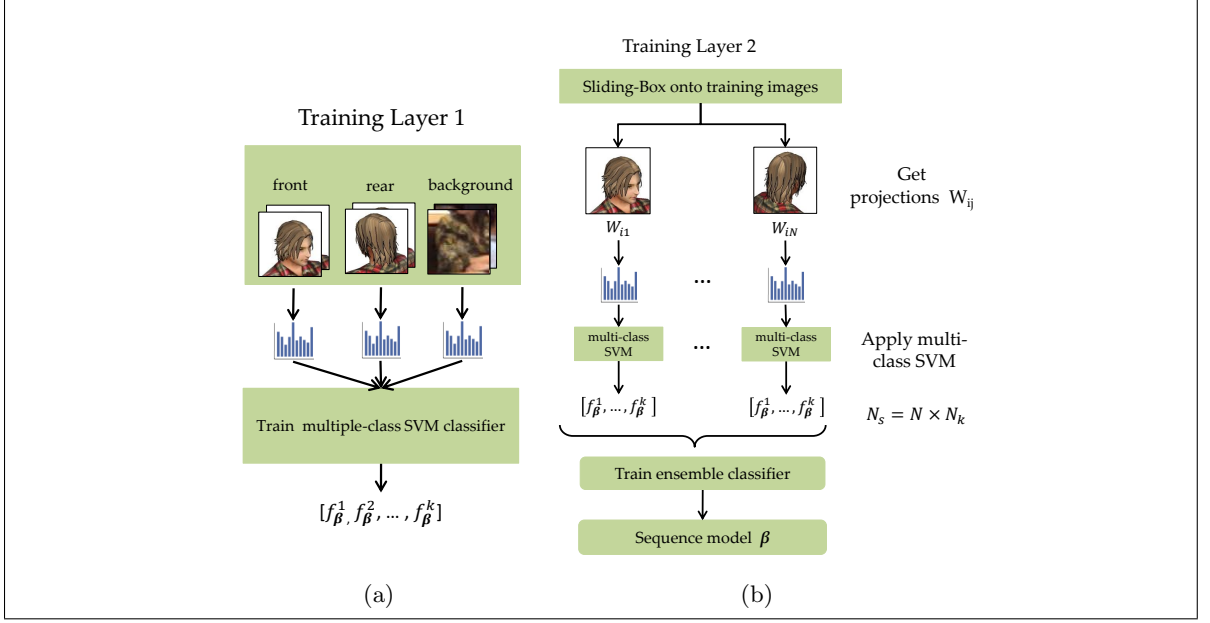


Figure 7: Training process diagram for the classifier ensemble scheme. We use an ensemble of classifiers divided into two layers: (a) Shows the first layer formed by a multi-class SVM classifiers. This layer identifies frontal head, backward head, and background. (b) Shows the second layer, which is trained using the scores $(f_{\beta}^1, \dots, f_{\beta}^k)$ obtained by applying the first layer of classifiers to each element W_{ij} . This process yielded the model β , which can classify the image sequence.

have a priori knowledge about the head aspect in the scene. To address this problem, we generate four collections of aspects from the input collection applying circular shifts. These shifts allow us to find the most confident collection of aspects according to the learned models. We apply the multiple view classifier to each collection and we choose the one with highest confidence, as shown Fig. 8. We do not use random shifts or a random combination of aspects, because an admissible collection of head aspects presents a coherent sequence of appearance according to our camera system, *e.g.*, the collection of aspects *backward – front – backward – front* is not allowed.

3.5 Non Maximal Suppression

Non-Maximal Suppression (NMS) is a critical procedure in computer vision algorithms in which one must choose the most representative detection from a set of close confident detections. We implement NMS applying two algorithms: a Weighted Mean Shift (WMS) (Dalal, 2006) and a Local Maximum Searching (LMS) (Dollar et al., 2009) in the 3D domain. The LMS algorithm is an heuristic search that chooses a final detection by applying a strategy which suppresses the less confident of every pair of detections with a significant overlap. This procedure assigns to the sliding-box \mathbf{B}_i the maximum score f_{β} found in this set of overlapped detections. We define the overlap criterion in terms of the Euclidian distance between the sliding-boxes.

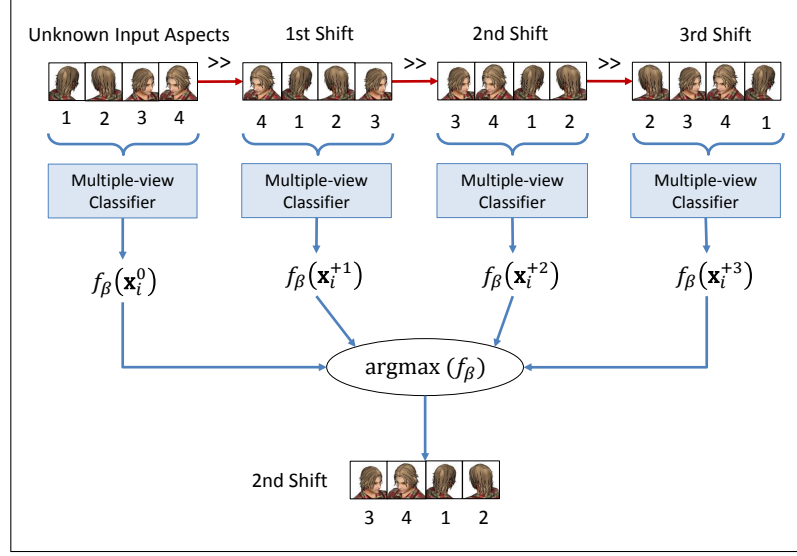


Figure 8: Diagram of best collection searching. The algorithm receives an input collection of aspects without a priori knowledge about its correct alignment. We apply a set of circular shifts to generate the total number of four collections of aspects, including the input collection. After applying the multiple view classifier to each collection, we choose the most confident one using an *argmax* criterion. In the diagram, the multiple view classifier assigns a set of confidence values to each collection of aspects: $f_{\beta}(\mathbf{x}_i^0) = -1.25$, $f_{\beta}(\mathbf{x}_i^{+1}) = -0.12$, $f_{\beta}(\mathbf{x}_i^{+2}) = 1.75$, $f_{\beta}(\mathbf{x}_i^0) = 0.75$. Finally, our algorithm selects the collection of aspects generated with the second shift because it best matches the training samples.

The WMS algorithm computes the local modes from a set of multiple detections. Each detection \mathbf{B}_i has an associated symmetric positive definite 3×3 bandwidth covariance matrix to define the smoothing width for the detected position \mathbf{M}_i . Overlapped detections are fused to represent the n points as local modes. The derivation is the same as in Dalal (2006). In our approach, we assume diagonal covariance matrices only considering the uncertainty of location because our sliding-box always has a fixed size and shape.

4 Datasets

We deploy a full calibrated and synchronized multiple view camera system in an indoor environment in order to capture labeled datasets. Specifically, we install cameras in the four upper corners of a classroom, as shown in Fig. 11b. We generate two full labeled multiple view datasets. We use one of these datasets for training and one for testing. Figs 9 and 11 show examples of these datasets, respectively. Notice that both datasets are captured under different conditions, so we can really test the generalization capabilities of the proposed method. Both datasets are acquired at 640×480 pixels and 15 fps.

Training dataset contains images in which a set of ten people are placed at six locations within a classroom, as shown in Fig. 9. People turn around their Z axis from 0° to 360° and we manually label head instances. We also use the mirrored versions of these labels

Table 1: Details of the training dataset used to train the feature ensemble. Each instance is a collection of aspects, as shown in Fig. 10.

class	number of examples
head	1,000
background	4,570
total	5,570

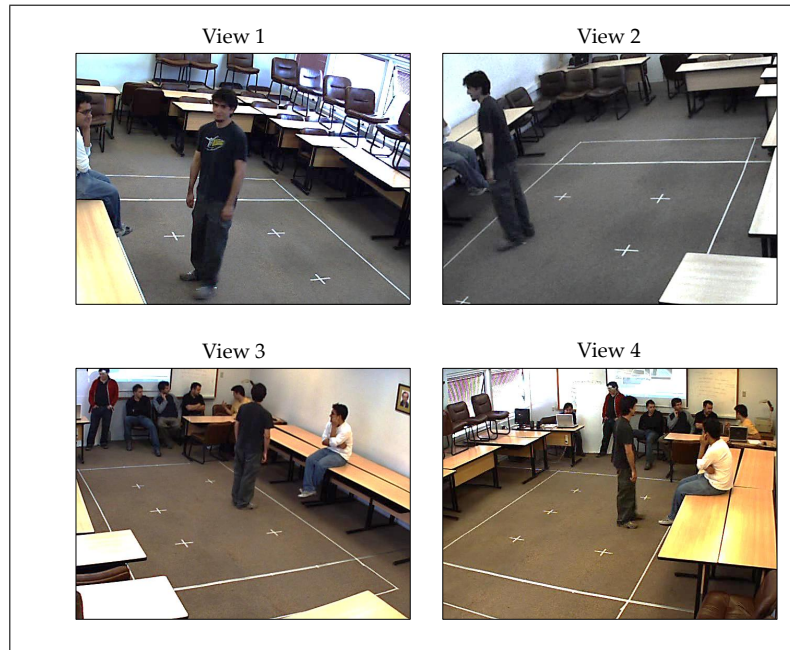


Figure 9: Example of images collected for training. Each image comes from one view in the camera system. To facilitate the labeling process, people stand on the white crosses on the floor and turn around their Z axis generating various views of the head (best viewed in color).

to generate new samples and enrich the training dataset. Positive training instances are collections of aspects of the sliding-boxes projected over the people’s heads and negative instances are sliding-box projections from random locations in the room and body parts, as shown in Fig. 10. Table 1 presents a summary of the training dataset. We use this dataset to train both classification schemes presented in Section 3.4, i.e., feature ensemble method and the second layer of the ensemble of classifiers. We also collect instances W_{ij} from the same dataset in order to generate a dataset of 16,494 single view examples separated into three classes: frontal, backward, and background, as shown in Table 2. Although we consider two individual aspects, frontal and backward, our approach is not limited to the number of poses for individual aspects and collections of aspects.

Test dataset consists of two fully labeled multi-view video sequences acquired in a classroom in which subjects present different activity levels. Both test sequences contain views of heads and torsos. An instance is defined as a detection when it appears in all cameras simultaneously. We test our algorithm over an average of 6,500 test instances. This number



Figure 10: Examples of aspects collection used for training. (a) positive instance of people’s head retrieved from multi-view camera system and (b) background samples of classroom environment. We use four cameras in both cases (best viewed in color).

Table 2: Details of the training dataset used to train individual models.

class	number of examples
Frontal head	2,000
Rear head	2,000
Background	12,494
total	16,494

of instances represents an average number of 26,000 labeled instances taking into account the four cameras. Table 3 shows a summary of both test sequences.

5 Experiments and Results

In this section, we describe the experimental results of applying the sliding-box approach to head detection. We present qualitative and quantitative results of our method, and compare our results to those of a state-of-art 2D multi-scale detector. In our implementation, we apply a dense grid mesh with 5cm steps to apply the sliding-box. We use Local Binary Patterns (LBP) as our appearance descriptor to quantify shape and texture cues Ojala et al. (2000). Following Bosch et al. (2007), we compute LBPs over a 3-level pyramidal decomposition, obtaining a total of 21 feature blocks per object instance. The final LBP descriptor contains $K = 59$ bins. We use the VLFeat library to compute the LBP feature Vedaldi and Fulkerson (2010). In our experiments, we use linear kernel to implement the multiple view classifiers. Our current implementation takes an average of six minutes to classify 1,500 boxes using MATLAB on Ubuntu-Linux and an AMD Phenom II X4-925, 4GB RAM, 2.8GHz computer. Besides the time execution limitations of MATLAB, the

Table 3: Details of the dataset used to test the proposed approach. Sequences are called *sq-01* and *sq-02*. We show the average number of people per image and the total number of frames in each sequence.

sequence	avg. no. people/image	no. of frames	avg. no. labels	avg. no. ground-truth instances
sq-01	7	245	6,000	1,700
sq-02	8	600	20,000	5,000

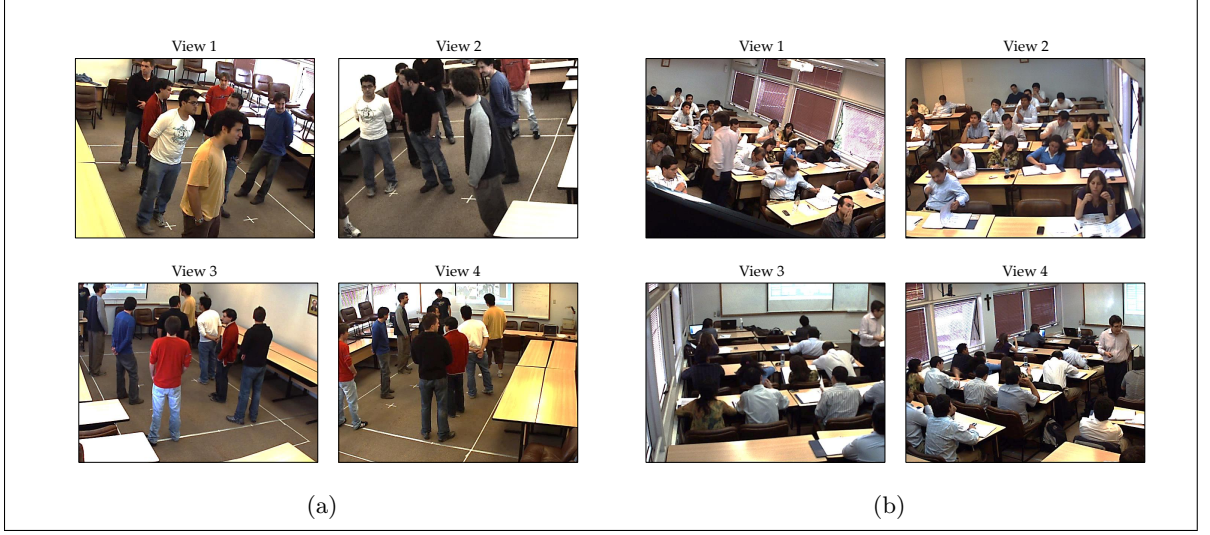


Figure 11: Example of images collected for testing. Both test datasets contain images of people in a classroom at different activity levels (best viewed in color). (a) The first test sequence contains people moving and changing their appearances. (b) The second test sequence consists of people sitting, observing a lecture.

sliding-box process can be easily implemented in parallel.

5.1 Detection Performance

In this section we evaluate the performance of our approach using the methodology suggested by the PASCAL visual recognition challenge Everingham et al. (2006). As a 2D baseline, we compare the performance of our approach with respect to the popular Deformable Part-based Model (DPM) proposed in Felzenszwalb et al. (2010b). We train the DPM using our head dataset. In order to obtain a fair comparison, we use epipolar geometry to filter out false positive detections of the DPM detector, as we explain below.

First, we use a sliding-window approach to execute the DPM on each camera view independently. Positive detections in each view are then verified in the other cameras using epipolar lines. We accept as a true detection in camera i only a detection that can be tracked through N_{match} views of the camera system. In our experiments, we use a conservative value of $N_{match} = 2$ to avoid filtering out valid detections. A detection in camera i is considered valid in camera j , if there is a positive detection near the corresponding epipolar line in camera j , such that $\mathbf{l}_j = \mathbf{F}_{ij} \cdot \mathbf{m}_i$, where \mathbf{F}_{ij} is the fundamental matrix between views i and j , and \mathbf{m}_i is the centroid of the detection in camera i . We also include object size and position priors in the sliding-window scheme by limiting the scales and image positions to select candidate windows.

Since our sliding box method operates in world coordinates, we slightly modify the Pascal criterion by evaluating performance using world coordinates instead of image coordinates.

Table 4: Detection performance in terms of Average Precision (AP) in sequences *sq-01* and *sq-02*. We report our methods EF, OVA-EC and OVO-EC. In all cases, we compare the performance using Local Maximum Searching (LMS) and Weighted Mean-Shift (WMS) as NMS procedures. Additionally, as a baseline, we include the performance of DPM and DPM using epipolar geometry.

Algorithm	Average Precision (AP)			
	LMS		WMS	
	sq-01	sq-02	sq-01	sq-02
EF	0.88	0.72	0.89	0.73
OVA-EC	0.83	0.72	0.82	0.73
OVO-EC	0.89	0.73	0.95	0.74
DPM	0.92	0.62	0.92	0.62
DPM-epipolar	0.90	0.57	0.82	0.57

Specifically, we modify the Pascal evaluation function by calculating Euclidian distance between a sliding-box detection SB_{dt} and the ground truth position SB_{gt} . Both, SB_{gt} and SB_{dt} are considered as a valid match if their distance is lower than the overlap radius $r_{overlap}$, which depends on prior information. In our experiments, we consider $r_{overlap} = 180[mm]$. This value is the average radius producing an average overlap of 50% for 2D detection according to the PASCAL criterion. For evaluation purposes, we restrict the region of interest S to scene areas where the field of view of the available cameras overlaps. For DPM and its version using the epipolar filter, we report the detection performance of the view with the highest AP value.

Table 4 shows that the OVO-EC model using WMS suppression outperforms the EF and the OVA-EC models. These results also show that the suppression procedure affects the overall detection performance. These variations are more clear in dataset *sq-01*, where suppression and head pose changes present a higher influence on detection performance. WMS suppression represents the overlapping boxes using a weighted mode that presents a higher fit with respect to ground-truth locations. In general, in terms of detection accuracy, the proposed approach outperforms the DPM detector, with the exception of *sq-01* dataset using LMS procedure. This dataset is acquired using the same environment configuration and part of the people in the training set. These similarities may explain the highest performance of the DPM detector. Both versions of the DPM detector present similar performance.

We compute the confidence interval for detection in order to access the repeatability and accuracy of the OVO-EC method using WMS suppression. This procedure takes k -folds of test instances and computes the AP value for each fold. Then it uses these k AP values to compute the average and the confidence intervals using a t -Student Test at 95% of confidence level. We set $k = 10$, which is the standard value used in cross-validation evaluations. Table 5 shows little variation in AP values for the OVO-EC method, proving the repeatability of our method. As expected, this test also indicates average performance close to the AP values presented previously for overall detection performance. Most intervals present a significance of over 95% confidence level and error below 0.05.

Table 5: Repeatability analysis of OVO-CE based on WMS suppression in dataset *sq-01* and *sq-02*. Results show little variation after detection and a significance over 95% of confidence level.

Method	<i>sq-01</i>			<i>sq-02</i>		
	Avg. AP	σ	Confidence Interval	Avg. AP	σ	Confidence Interval
OVO-EC	0.953	0.025	[0.935 - 0.971]	0.754	0.034	[0.730 - 0.778]
OVA-EC	0.855	0.058	[0.813 - 0.896]	0.855	0.058	[0.813 - 0.896]
EF	0.926	0.037	[0.899 - 0.952]	0.926	0.037	[0.899 - 0.952]

Figure 12 shows a comparison between detection provided by the OVO-CE method based on WMS suppression and the DPM detector. We consider the case of the view where the DPM presents the highest AP value. This result shows the improvements achieved by our method in retrieving detections missed by DPM. We use red boxes to indicate detections of our method and green boxes to indicate ground-true positions of heads. In general, our approach successfully detects heads that are not detected by the DPM model. Hereafter, we focus on the OVO-CE method because it provides the best performance.

5.2 Enriched features

We analyze the influence of training the multiple view classifier varying the number of cameras, in order to evaluate the impact of using different visual sources and the power of combining features from various view-points of the same object. First we train the model using only one camera. Then we progressively re-train the model adding one camera at a time. We perform the analysis in terms of per-window performance using AP value without considering the suppression procedure after classification. Figure 13 shows an increase in performance as we add a new camera. This confirms our claim that if we use more cameras, we enrich the features set with information provided from different viewpoints. As expected, we achieve the highest performances when the classifier uses the four views. When one camera is used, we report a low AP rate, showing that the classifier using one viewpoint generates weak models. We also note that the models intrinsically perform the task of aligning the collection of aspects that they learnt during training. Figure 14 shows two sets of four candidate boxes. The first set shown in Fig.14a, belong to the head class; the second set, in Fig.14b, belongs to the background class. During the evaluation of each hypothesis, higher scores are always given to the best alignments. All of the scores in the second set b) are strongly negative, and therefore they are assigned to the background class.

5.3 Focus-of-Attention

The focus-of-attention procedure filters detections to decrease the computational burden during detection. This procedure uses the output of an interest-point detector to estimate the spatial position of sliding-box candidates by triangulation Hartley and Zisserman (2003). In our implementation, we use the detections provided by a 2D detector that works as a



Figure 12: Comparison between detections of DPM and OVO-EC method using WMS suppression. Results show that OVO-EC is able to correctly detect heads that are not detected by the DPM model. We use red boxes to indicate detections of our method and green boxes to indicate ground-true positions of heads (best viewed in color). (a) and (c) detections generated by DPM detector. (b) and (d) detections generated by OVO-EC using WMS suppression.

specialized interest-point detector. Figure 15 shows that the total number of hypotheses and therefore the maximum recall that we can achieve depends on detector confidence, *i.e.*, the higher the confidence threshold, the smaller the number of spatial hypotheses and the chances of recovering the total number of heads. Table 6 summarizes the burden reduction by using focus-of-attention. The average burden reduction is 95% at a threshold that allows us to recover all detections in the region of interest S . Although there is an overhead due to applying an interest-point detector, and a trade-off between the number of spatial hypotheses and the detector confidence, our experiments indicate that the proposed focus-of-attention mechanism helps to significantly reduce the computational complexity.

Figure 16 shows a heat map visualization that summarizes the locations of heads projected onto the ground plane. We compute these heat maps by the cumulative log-normalized sum of the detections in all frames in the sequences. The highest hits in red denote the

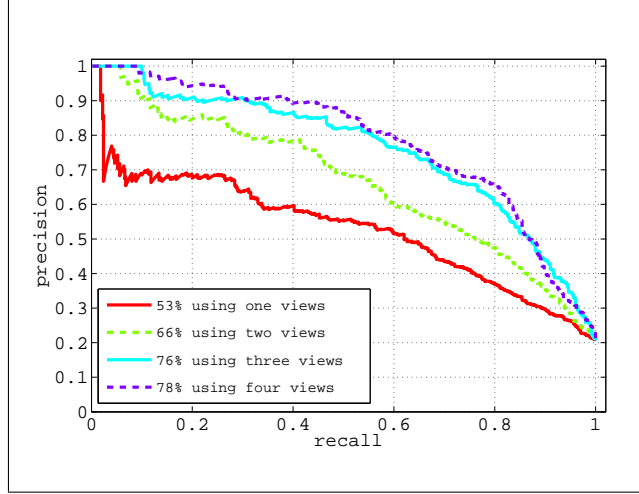


Figure 13: Performance comparison in terms of the cameras used for the OVO-CE method. The curves show performance evolution as we add information from the four visual sources of our camera system. Overall performance is represented using the AP value. Higher AP values are obtained as we add more views to the ensemble. We conclude that the additional viewpoints help to improve performance.

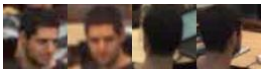
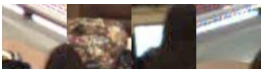
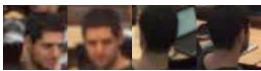
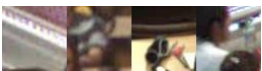
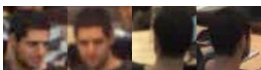

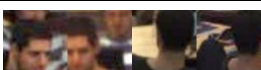
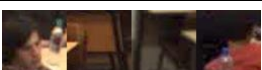
Table 6: Summary of the burden reduction due to the spatial focus-of-attention. Although there is an overhead due to applying an interest-point detector, the proposed focus-of-attention mechanism helps to significantly reduce the computational complexity.

Sequence	Total no. of Boxes	Percentage of Burden Reduction Varying the Detector Threshold						
		-1.0	-0.7	-0.4	0	0.4	0.7	1.0
<i>sq-01</i>	92,160	85 ± 3.5	95 ± 1.5	98 ± 0.65	99 ± 0.5	99.5 ± 0.3	99.7 ± 0.27	100 ± 0.15
<i>sq-02</i>	65,536	92.5 ± 3	96.3 ± 1.6	98.3 ± 0.4	99.3 ± 0.5	99.4 ± 0.5	99.6 ± 0.3	99 ± 0.2

location that contains the highest density of detections. Areas in blue correspond to those with few or no detections selected as head candidates. In both heat-maps, we observe the effectiveness of the focus-of-attention to produce spatial hypotheses around the ground-truth and simultaneously reject large areas close to the edges of S , as shown in Fig. 16c and Fig. 16d.

5.4 Using Geometrical Cues to Select Candidate Locations

We perform a qualitative evaluation comparing the windows used by sliding-window detectors with our multiple view detector. We select two sets from a random sample of 60 windows in both detection approaches. Windows used with the 2D detector cover various scales trying to predict the size of the object, which produces false positives at scales with higher resolutions, Fig. 17a. Our approach can control the size of the windows and the areas analyzed. This allows us to generate more informative windows for the classifier and limit the analysis to more likely location, as shown in Fig. 17b.

Image Example	Score	Image Example	Score
	0.74		-2.20
	1.24		-3.28
	1.66		-4.54
	-1.92		-3.83

(a) (b)

Figure 14: Alignment test of collection of aspects. Figure shows the results of using sliding-box to evaluate among a set of hypotheses and their scores. Box scores are higher when the box belongs to the head class, and its projections yield better alignment, as shown in (a). In (b) we observe background examples and their scores, all of which are negative (best viewed in color).

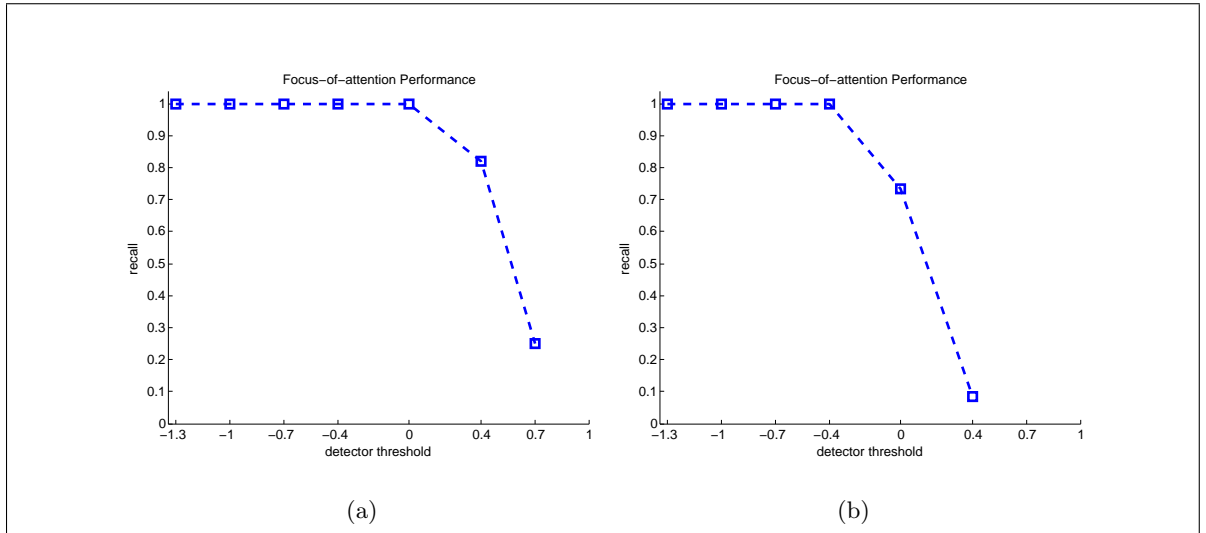


Figure 15: Sensitivity analysis of confidence vs recall due to use of the spatial focus-of-attention procedure. There is a trade-off between detector sensitivity and the maximum recall reached upstream of the classifier. If we set a low sensitivity for the detector, we increase the ability of the focus-of-attention to retrieve all detections in the scene. (a) Shows sensitivity for sequence *sq-01* where the maximum recall occurs between 0 and -0.4 of detector confidence. (b) Shows sensitivity for sequence *sq-02* where the maximum recall occurs between -0.4 and -0.7 of detector confidence.

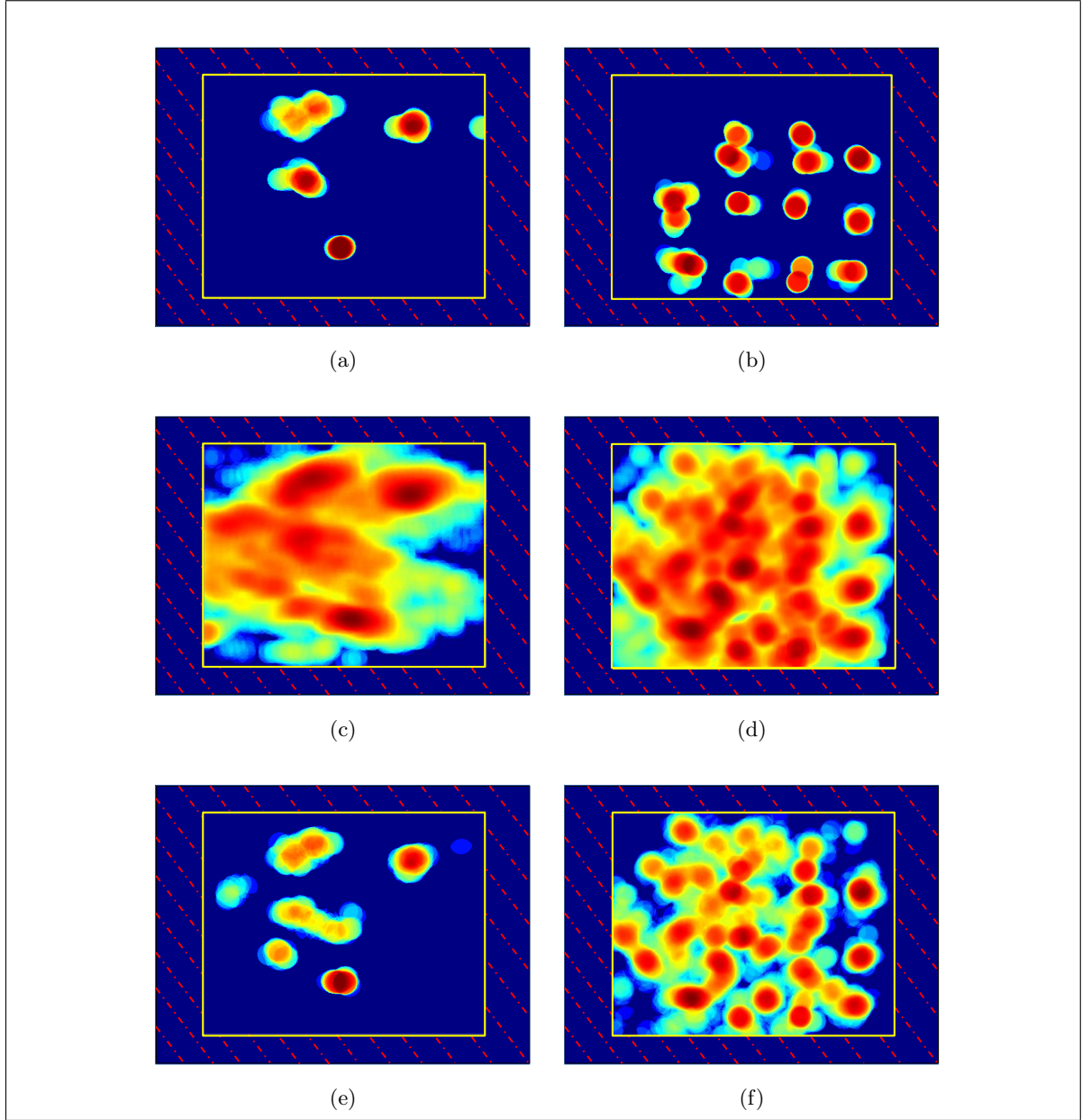


Figure 16: (a) Heat map representations of the ground-truth and focus-of-attention for both sequences. (b) Heat maps for ground-truth heads in the test sequences. (c) and (d) outputs after applying the focus-of-attention procedure. This produces a reduced set of spatial hypotheses that allow us to estimate regions close to real heads. We also use geometric priors to limit the analysis to the space defined by the yellow square S , and to discard the crosshatch area that does not contain feasible candidate points. Note the discarded areas close to the edges of S in both sequences that are perceived easily in sequence *sq-01*, as shown in (c).

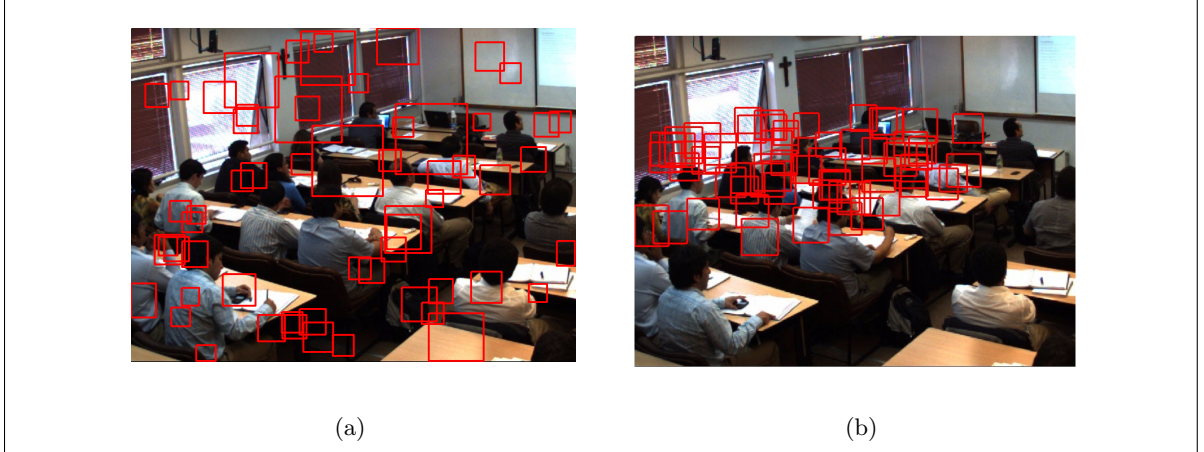


Figure 17: Random sample of 60 windows evaluated: (a) by a 2D detector, and (b) by our approach. In (a), windows must change their size in order to predict the real size of the object. We improve the search by using the real size of the object to limit the size and locations of the projected windows, as shown in (b).

6 Conclusions

We describe a sliding-box approach to detect heads in indoor multiple view environments. This method is an extension of the sliding-window approach to 3D spaces. It allows information from various viewpoints to be combined based on their corresponding image regions, thus enabling us to recover spatial information during detection. With this strategy we outperform single view detection methods described in the state-of-the-art.

Results show that OVO-EC based on WMS generally outperforms other methods in terms of AP. We conclude that the ensemble of classifiers approach allows us to represent each projection of the sliding-box in the collection of aspects using a compact representation. Therefore, we can imagine the first layer of the ensemble as a group of mid-level features. The second layer performs the mixture of these parts. Although detection performance may depend on classification schemes, we found that the main differences are due to the suppression method used to eliminate overlapping detections. Mean-shift suppression improved these performances in both datasets. Experiments also show the ability of the sliding-box to integrate detections provided by a single view detector using an integrated multiple view feature descriptor. This makes it possible to recover missed detections generated by a 2D detector because the integrated descriptor presents higher robustness to pose variations. Additionally, our results indicate that the proposed focus-of-attention mechanism as a pre-filter step helps to further restrict the search space and drastically reduce the number of candidate hypotheses.

Current limitations to our method are the calibration process used to recover the spatial information from the scene, which makes our approach somewhat rigid to the scene structure. Also, the projection process presents a high running time that prevents us from

testing large datasets. Future work will focus on extending our method to semi-calibrated or uncalibrated multiple view configurations. We also plan to address computational issues exploring parallel implementation and the use of hashing techniques.

Acknowledgment

Part of our research was funded by project ACT-32, project FONDECYT N.1100830, and FONDEF grant associated to the project D10I1054. We also thank Pablo Espinace for his help in beginning of this research..

References

- Ali, I. and Dailey, M. (2012). Multiple human tracking in high-density crowds. *Image and Vision Computing*.
- Ali, I. and Dailey, M. N. (2009). Multiple human tracking in high-density crowds. In *Advanced Concepts for Intelligent Vision Systems (ACIVS)*, pages 540–549. Springer.
- Bosch, A., Zisserman, A., and Muñoz, X. (2007). Image Classification using Random Forests and Ferns. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8. IEEE.
- Chang, R., Chua, T. W., Leman, K., Wang, H. L., and Zhang, J. (2013). Automatic cooperative camera system for real-time bag detection in visual surveillance. In *Proceedings of the IEEE International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–6. IEEE.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.
- Cyr, C. M. and Kimia, B. B. (2001). 3d object recognition using shape similarity-based aspect graph. In *IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 254–261. IEEE.
- Cyr, C. M. and Kimia, B. B. (2004). A similarity-based aspect-graph approach to 3d object recognition. *International Journal of Computer Vision*, 57(1):5–22.
- Dalal, N. (2006). *Finding people in images and videos*. PhD thesis, Institut National Polytechnique de Grenoble-INPG.

- Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE.
- Dean, T., Ruzon, M. A., Segal, M., Shlens, J., Vijayanarasimhan, S., and Yagnik, J. (2013). Fast, accurate detection of 100,000 object classes on a single machine. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1814–1821. IEEE.
- Delannay, D., Danhier, N., and De Vleeschouwer, C. (2009). Detection and recognition of sports (wo) men from multiple views. In *ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–7. IEEE.
- Dollár, P., Belongie, S., and Perona, P. (2010). The Fastest Pedestrian Detector in the West. In *British Machine Vision Conference (BMVC)*.
- Dollar, P., Tu, Z., Perona, P., and Belongie, S. (2009). Integral Channel Features. *British Machine Vision Conference (BMVC)*.
- Dollar, P., Wojek, C., Schiele, B., and Perona, P. (2011). Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, pages 1–20.
- Eshel, R. and Moses, Y. (2010). Tracking in a dense crowd using multiple cameras. *International Journal of Computer Vision*, 88(1):129–143.
- Espinace, P., Kollar, T., Roy, N., and Soto, A. (2013). Indoor scene recognition by a mobile robot through adaptive object detection. *Robotics and Autonomous Systems*, 61(9):932–947.
- Everingham, M., Zisserman, A., Williams, C., and Van Gool, L. (2006). The PASCAL Visual Object Classes Challenge 2006 (VOC2006) results.
- Felzenszwalb, P., Girshick, R., and McAllester, D. (2010a). Cascade object detection with deformable part models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2241–2248. IEEE.
- Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010b). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence (PAMI), 32(9):1627–1645.

Girshick, R. B., Donahue, J., Darrell, T., and Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. *Computer Research Repository (CoRR)*, abs/1311.2524.

Hartley, R. and Zisserman, A. (2003). *Multiple View Geometry in Computer Vision*, volume 2. Cambridge University Press.

Hayashi, M., Yamamoto, T., Aoki, Y., Ohshima, K., and Tanabiki, M. (2013). Head and upper body pose estimation in team sport videos. In *Proceedings of the IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 754–759. IEEE.

Kim, K. and Davis, L. (2006). Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 98–109.

Koenderink, J. J. and van Doorn, A. J. (1979). The internal representation of solid shape with respect to vision. *Biological cybernetics*, 32(4):211–216.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 2169–2178. IEEE.

Liem, M. C. and Gavrilu, D. M. (2013). A comparative study on multi-person tracking using overlapping cameras. In *Proceedings of the International Conference on Computer Vision Systems (ICVS)*, pages 203–212. Springer.

Mundy, J. L. (2006). Object recognition in the geometric era: A retrospective. In *Toward category-level object recognition*, pages 3–28. Springer.

Nghiem, A. T., Auvinet, E., and Meunier, J. (2012). Head detection using kinect camera and its application to fall detection. In *Proceedings of the International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pages 164–169. IEEE.

Ojala, T., Pietikäinen, M., and Mäenpää, T. (2000). Gray scale and rotation invariant texture classification with local binary patterns. In *Proceedings of the European Conference*

on *Computer Vision (ECCV)*, pages 404–420. Springer.

Pane, C., Gasparini, M., Prati, A., Gualdi, G., and Cucchiara, R. (2013). A people counting system for business analytics. In *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 135–140. IEEE.

Papageorgiou, C. and Poggio, T. (2000). A Trainable System for Object Detection. *International Journal of Computer Vision*, 38(1):15–33.

Pedersoli, M., González, J., Bagdanov, A., and Villanueva, J. (2010). Recursive coarse-to-fine localization for fast object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 280–293. Springer.

Pedersoli, M., Gonzalez, J., Hu, X., and Roca, X. (2014). Toward real-time pedestrian detection based on a deformable template model. *IEEE Transactions on Intelligent Transportation Systems*, 15(1):355–364.

Salas, J. and Tomasi, C. (2011). People detection using color and depth images. In *Pattern Recognition*, pages 127–135. Springer.

Savarese, S. and Fei-Fei, L. (2007). 3D generic object categorization, localization and pose estimation. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1–8.

Spinello, L. and Arras, K. O. (2011). People detection in rgb-d data. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3838–3843. IEEE.

Stone, J. V. (1999). Object recognition: View-specificity and motion-specificity. *Vision Research*, 39(24):4032–4044.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer.

Tang, S., Andriluka, M., Milan, A., Schindler, K., Roth, S., and Schiele, B. (2013). Learning people detectors for tracking in crowded scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1049–1056. IEEE.

Tarr, M. J. and Kriegman, D. J. (2001). What defines a view? *Vision Research*, 41(15):1981–

2004.

Vedaldi, A. and Fulkerson, B. (2010). Vlfeat: An open and portable library of computer vision algorithms. In *Proceedings of the International Conference on Multimedia*, pages 1469–1472. ACM.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–511. IEEE.

Viola, P., Jones, M., and Snow, D. (2005). Detecting pedestrians using patterns of motion and appearance. *International Journal of Computer Vision*, 63(2):153–161.

Xie, D., Dang, L., and Tong, R. (2012). Video based head detection and tracking surveillance system. In *Proceedings of the International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 2832–2836. IEEE.

Yang, J., Yu, K., Gong, Y., and Huang, T. (2009). Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1794–1801. IEEE.

Zeng, C. and Ma, H. (2010). Robust head-shoulder detection by pca-based multilevel hog-lbp detector for people counting. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, pages 2069–2072. IEEE.