NATIONAL UNIVERSITY OF SCIENCE AND TECHNOLOGY POLITEHNICA BUCHAREST

Doctoral School of Electronics, Telecommunications and Information Technology

Decision No. 149 **from** 23-11-2023

Ph.D. THESIS SUMMARY

Eng. Andrei - Cosmin JITARU

TEHNICI DE ÎNVĂȚARE AUTOMATĂ PENTRU ANALIZA VIZUALĂ A INTERACȚIUNII UMANE

MACHINE LEARNING - BASED VISUAL ANALYSIS OF HUMAN-TO-HUMAN INTERACTION

THESIS COMMITTEE

Prof. dr. ing. Mihai CIUC National Univ. of Science and Technology Politehnica Bucharest	President
Prof. dr. ing. Bogdan IONESCU National Univ. of Science and Technology Politehnica Bucharest	PhD Supervisor
Prof. dr. ing. Ruxandra ȚAPU National Univ. of Science and Technology Politehnica Bucharest	Referee
CS II dr. ing. Adrian POPESCU CEA-LIST, France	Referee
Conf. dr. ing. Ioan BUCIU University of Oradea	Referee

BUCHAREST 2023

The thesis has been funded by the Doctoral scholarship financed by the Ministry of National Education and partly funded by the Ministry of Investments and European Projects through the Human Capital Sectoral Operational Program 2014-2020, Contract no. 62461/03.06.2022, SMIS code 153735.

Table of contents

Part I

Introduction and Theoretical Aspect

Chapter 1

Introduction

1.1 Domain of the thesis

In contemporary discourse within academia and technology-focused media, the subject of public safety has garnered significant attention. Concomitant with the rapid evolution of digital transformation paradigms, the emergence of big data phenomena pervades various sectors, ranging from retail environments to public transportation and open public spaces. Within this context, visual sensing mechanisms, including video surveillance cameras and imagery systems, necessitate increasingly sophisticated data processing frameworks. These frameworks aim to convert the voluminous raw sensor data into actionable insights through the application of advanced data analytics methodologies.

The availability of huge data for analysis complicates AI trainable data topics. The majority of collecting solutions lack metadata for AI analysis tools. In-the-wild data is essential for translating scientific findings into production. Going further, one more aspect which makes the research field of the thesis a niche is the comprehension of the human-to-human interaction in the surveillance domain of computer vision techniques.

In the context of human-to-human interactions, extending insights from both analytical and operational engineering perspectives to cultivate machine learning-friendly datasets, the anticipated deliverables from analytics tools at the decision-making level encompass: streamlining the automatic interpretability of a given scene, vvaluation of data from heterogeneous sources, facilitate the deployment of AI-driven algorithms for data configurations that are too intricate for human-centric analysis and automatize procedural processes in the development of decision-making frameworks.

1.2 Motivation of the thesis

The real of machine learning, particularly in visual analysis, has seen exponential growth in applications. The successful execution of a machine learning task is not just reliant on an appropriate dataset, but also the intuitive comprehension of the human operator towards the problem at hand.

The field of deep learning has made significant advancements in the domain of voice recognition. However, despite these improvements, certain issues such as overlapping speech and noise continue to pose obstacles in achieving optimal performance. The application of visual speech recognition in the context of the Romanian language is subject to some restrictions, which include the scarcity of available datasets and the presence of distinct lip movements specific to this language. Advanced recognition is very advantageous for remote sensor surveillance, particularly in expansive regions. The present study explores the application of visual analytics in the context of group dynamics, with a specific focus on its potential to enhance public safety during large-scale events. Furthermore, the study of human-media interaction, encompassing the generation of synthetic material and the impact on media memorability, is gaining increasing scientific and societal importance as digital content continues to advance.

These aspects have driven my interest in exploring the aforementioned research areas. In this thesis, I delineate details about data collection and dataset development for real-world scenarios. Additionally, several methodologies for crafting systems with integrated deep learning techniques suitable for unconstrained environments are presented. This discussion is complemented by best practices for enhancing system efficiency and invaluable insights gathered for prospective advancements.

1.3 Content of the thesis

The thesis is structured as follows: Chapter 2 focuses on the correlation between image analysis and deep learning, with a specific emphasis on theoretical factors, AI trainable data development and the main challenges that impede the transferability of deep learning to real-world scenarios. Part II describes my contributions to machine learning, computer vision datasets, and real-world analytics systems. Chapter 3 discusses visual speech recognition. Chapter 4 describes a full system for aerial imaging task analysis. Detailed components of coherent human crowd grouping dynamics are discussed to address complex difficulties. A unique algorithm for human crowd scene detection is shown, resulting in a still image collection for coherent crowd analysis. Chapter 5 covers media

memorability predictions and deepfake detection. The thesis concludes with Chapter 6, which synthesizes the original contributions and suggests future research.

Chapter 2

Human-to-Human Interaction

This chapter covers theoretical aspects, key neural network architectures for image analysis, in-the-wild AI data gathering, performance optimization with custom datasets and challenges in deploying in-the-wild deep learning.

2.1 Foundational Principles

2.1.1 Speech recognition

In human communication, auditory and visual modalities, encompassing voice and visual cues play pivotal roles. The process of identifying vocal cues is termed audio speech recognition (ASR), while the decoding of visual cues is referred to as visual speech recognition (VSR).

A VSR process examines an individual speaker's lip movement to identify spoken words in the Romanian language commencing with unprocessed video recordings and then a manual slicing procedure is essential to supply the deep learning model with a frame sequence.

2.1.2 Aerial monitoring

An imagery product, such as a raster, typically covers areas exceeding 20,000 x 20,000 pixels, necessitating substantial processing effort. Hence, a preprocessing slicing algorithm is essential. The following equations detail the overlap factor (Equation 2.1) and the validation window factor (Equation 2.2).

$$overlap_{fact} = \frac{\underset{o<i<N_t}{max}(max(L_{t_i})), \underset{o<i<N_t}{max}(max(l_{t_i}))}{GSD \cdot max(H,W)} \tag{2.1}$$

$$zero_{fact} = \frac{max_{o<i<N_t}(max(L_{t_i})), max_{o<i<N_t}(max(l_{t_i}))}{GSD \cdot min(H,W)} \tag{2.2}$$

where L_t and l_t represent the length and width of objects in centimetres, N_t represents the total number of annotated objects, GSD represents the ground sample distance in centimetres % pixels and H and W represents the height and width of the image in pixels.

By integrating the $overlap_{fact}$ and $zero_{fact}$ into the preprocessing algorithm, a slicing window will exhibit four overlapping regions with its neighbouring windows.

2.1.3 Crowd analysis

When deep learning methods are employed for crowd analysis, the primary objective commonly pertains to crowd counting, wherein the singular output entails the estimation of the number of pedestrians present in the assessed scene. In conjunction with this task, certain algorithms specialized in other domains of computer vision can yield valuable insights for security teams. These insights may encompass the automatic labelling of scenes as either "crowded" or "uncrowded" and the automatic identification of crowds within static images, concurrently delineating the boundaries of coherent groups.

2.1.4 Human and media products interaction

The interaction between humans and media products can readily be translated into its impact on memorability and the effects of deepfakes on human comprehension of information in the context described above. Deepfakes are computer-generated images or videos that convincingly mimic the appearance and behaviour of real individuals, often used to manipulate or deceive viewers. This explores the intersection of memorability and the effects of deepfakes on human information understanding, shedding light on the cognitive processes involved and the potential consequences for society.

2.2 Deep learning for image analysis

Deep neural networks, with their intricate hidden layer structures and increased depth compared to traditional networks, excel in solving complex problems by capturing subtle nuances. This deep learning landscape includes various models, from conventional architectures like Convolutional Neural Network (CNNs) and Recurrent Neural Networks (RNNs) to advanced approaches like Long Short-Term Memory (LSTMs), all contributing to recent advancements in deep learning.

2.2.1 Convolutional Neural Networks

The CNN architecture is carefully designed to automatically learn spatial hierarchies of features through backpropagation. It utilizes key components like convolutional layers, pooling layers, and fully connected layers to excel in feature extraction and representation learning.

Convolution

Convolution serves as a specialized linear operation primarily employed for feature extraction. Two critical hyperparameters define the convolution operation: the "size" and the "number of kernels".

Pooling

Similar to the Convolutional Layer, the Pooling layer functions to reduce the spatial dimensions of the Convolved Feature.

Fully convolutional layer

In the realm of scientific research, it is customary to treat the output feature maps that arise from the final convolution or pooling layer as a flattening step. This process involves converting them into a one-dimensional (1D) array consisting of numerical values, generating a vector.

2.2.2 Recurrent Neural Networks

Within the domain of deep learning, algorithms such as Recurrent Neural Networks (RNNs) and Long Short-term Memory (LSTM) networks are specifically well-suited for managing sequential data types, including time series data, speech, and textual information. These algorithms possess the unique capability to preserve and retain contextual information and memory across time steps, enabling them to formulate predictions or make decisions that are informed by prior input data.

2.2.3 Evaluation metrics

Accurate assessment of machine learning models is crucial for their improvement. In this section, we'll explore relevant metrics for lip reading, classification, object detection and segmentation models, diving into the mathematics behind them.

Confusion matrix

In a classification task, the confusion matrix provides a tabular representation elucidating the model's accuracy by juxtaposing predicted outcomes with actual labels. This matrix enumerates true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) derived from the model's assessments.

Precision and Recall

Precision evaluates the model's accuracy by assessing the ratio of True Positives to all predicted positives. While, Recall focuses on quantity, gauging the model's capacity to identify True Positives from all actual positives. It's also referred to as sensitivity.

The Precision-Recall curve facilitates the identification of an optimal threshold to strike a balance between these metrics.

F-measure

The F-measure or F-score, is a metric calculated using Precision and Recall, based on both ground truth and algorithm predictions.

Intersection over Union

In object detection, the Intersection over Union (IoU) serves as a pivotal metric for appraising the congruence between the predicted and ground truth bounding boxes.

Kappa score

The Cohen's Kappa coefficient enhances model efficacy evaluation, particularly for imbalanced datasets. It accounts for class distribution and instance discrepancies, offering a comprehensive view of model performance.

Mean average precision @k

Mean Average Precision at k (mAP@k) evaluates object detection and image retrieval models, comparing detected and ground-truth bounding boxes, and assigning scores based on prediction accuracy.

Accuracy and Top-k

The Accuracy evaluation metric is one of the baseline metrics in the computer vision domain, evaluating in what proportion the model predicts correctly.

Mean Absolute Error and Root Mean Squared Error

Mean Absolute Error (MAE) quantifies the disparities between paired observations representing the same phenomenon by calculating the average of their absolute differences. Root Mean Squared Error (RMSE) utilizes a quadratic scoring rule, which means it penalizes larger errors more significantly.

Dice Similarity Coefficient

The Dice coefficient, often referred to as the "Sørensen–Dice coefficient", serves as a standard metric for determining the similarity between two sets.

Area Under Curve

The Receiver Operating Characteristic (ROC) curve visually depicts how binary classifiers perform across different classification thresholds, showing the trade-off between TP and FP rates.

2.3 Image analysis data development

2.3.1 Data gathering

In computer vision, data acquisition complexity results from various collection methods, not just data volume. Effective management involves task-specific filtering, metadata use for efficient searches, and centralized collection to enhance automation, supporting AI and data engineering.

2.3.2 Data preprocessing

Preprocessing is essential for preparing AI datasets, bridging raw data acquisition and model training through cleaning, transforming, and feature extraction. The use of deep learning tools for dataset formation and metadata labelling is crucial, as these methodologies significantly enhance data quality for AI training.

2.3.3 Data augmentation

Supervised machine learning and deep learning models require extensive datasets and utilize data augmentation, such as temporal frame insertion and visual transformations, to enhance generalization and reflect real-world variations.

2.4 Good practices for Deep Learning

2.4.1 Hyper-parameter tuning

In machine learning, hyperparameter tuning is crucial for refining a model's performance by meticulously adjusting settings like learning rate and batch size to minimize loss and errors. This fine-tuning process defines the boundaries within which the learning algorithm optimizes itself based on the input data, ultimately aiming to achieve the most accurate results possible.

2.4.2 Knowledge transfer and knowledge leveraging in machine learning

The primary objective of transfer learning is to maximize the transfer of knowledge from the model's prior training task to the new task at hand.

Knowledge leveraging across datasets entails the amalgamation of similar datasets, particularly when both tasks share identical input data.

2.5 Present challenges of deep learning in-the-wild

Deep learning in uncontrolled multimedia environments is hindered by data collection and preprocessing challenges. The demand for explainable AI and standardized datasets grows as traditional evaluation metrics falter against real-world variability. The advent of large-scale models amplifies concerns over computational demands and the need for automated optimization, driving the pursuit of models that emulate human-like continuous learning and adaptation.

2.6 Conclusions

The aforementioned concepts and theoretical illustrations demonstrate how artificial intelligence, specifically within the domain of computer vision, adeptly converts unprocessed, real-world data into practical knowledge by use of a supervised deep learning model, a digital instrument.

Part II

Personal Contributions

Chapter 3

First Romanian Lip Reading System

Chapter 3 focuses on short-term video analysis, specifically focusing on targeted keyword recognition within video segments. To establish a benchmark for under-resourced languages such as Romanian, Section 3.1 details the development methodology for two proprietary datasets, LRRo Lab and LRRo Wild, and introduces outline baseline algorithms. Section 3.2 offers a language-agnostic training approach designed to decode world-level messages across three different languages.

3.1 LRRo: A Lip Reading Data Set for the Under-Resourced Romanian Language

3.1.1 Introduction

Lip reading, also known as Visual Speech Recognition (VSR), is derived from visual-only recordings of a speaker's lips. This skill enables the comprehension or sensing of the conveyed message, whether by a human lip reader or a machine. While lip reading is present across all languages, existing VSR frameworks cannot employ transfer learning due to the unique viseme differences in each spoken language.

Recent efforts have expanded beyond English-only databases for lipreading, such as LRW [13] and LRS2 [12], others, GRID Corpus [7] or LRS3 [4], leading to the development of the LRW-1000 [64], a comprehensive Mandarin dataset.

Fig. 3.1 Representative speakers from LRRo datasets (Lab & Wild). [26]

The Romanian broadcasting sector offers few programs for deaf viewers. Our dataset, LRRo, aims to enhance Romanian Visual Lip-Reading (VLR) systems, with a focus on the medical and security sectors.

3.1.2 Proposed datasets

In this section, we delve into the structure and annotation of the proposed datasets. The creation of these datasets follows a multi-stage procedure that merges manual tagging with automated tools for refining the data. The LRRo dataset comprises two subsets: the Wild LRRo and the Lab LRRo. Both have been constructed to ensure adequate samples for a lip-reading system. This includes having a diverse array of speakers, varied speech rates, and multiple backgrounds, particularly aiming at lesser-studied languages like Romanian.

The collection includes over 1,200 minutes from TV shows and organic speech recordings. The raw footage for Wild LRRo was sourced from YouTube[1]. In contrast, Lab LRRo was captured in a controlled lab setting, with the speaker positioned in front of a camera. Each LRRo data instance is presented as mouth crops. A scene's relevance was determined by the presence of a face for a continuous 5 seconds, based on the concept of useful face appearance. The dataset showcases a broad spectrum of speaker appearances, including varying angles, lighting conditions, and makeup. Samples of this are illustrated in Figure 3.1.

Wild LRRo dataset

The raw footage was sourced and segmented from a range of open-source recordings, including Romanian TV shows on topics like IT and social matters, news segments covering politics, economics, and drama, as well as Romanian TEDx presentations.

The audio spectrogram was leveraged to streamline the annotation process and the Aegisub tool[2] was employed for annotation. The transcriptions provided by the 18 volunteer annotators underwent verification by two master annotators.

[1] www.youtube.com
[2] www.aegisub.org

Lab LRRo dataset

In the Lab LRRo dataset, we recorded participants from a frontal view and a 30° angle using a secondary camera. Nineteen mostly student participants read text from a prompter. Some faced reading or speaking difficulties, like dyslexia or rhotacism. We designed 74 sentences to ease reading from our lexicon. We annotated videos using two ASR systems, detailed in [20] and [21]. Their transcriptions were compared against our pre-defined speech text with a specific lexicon.

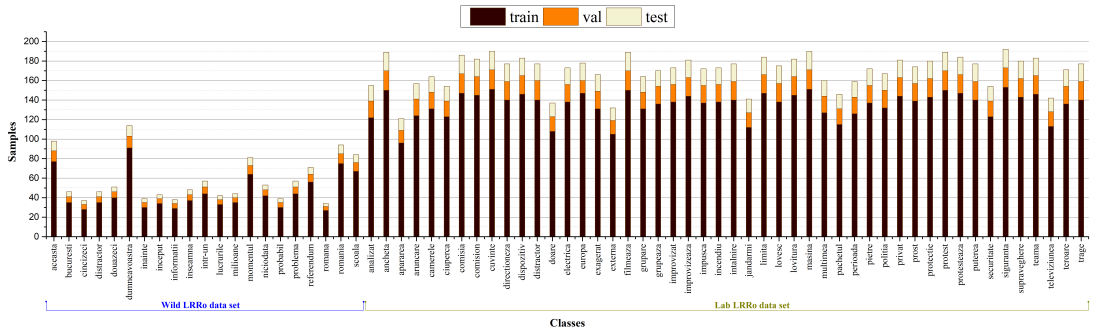


Fig. 3.2 Training, Validation, and Testing Split for the LRRo Datasets. [26]

3.1.3 LRRo dataset distribution

This section outlines the organizational structure of the publicly available LRRo dataset directories [26] [3].

The LRRo dataset, critical for Romanian language lipreading, comprises two distinct subsets: Wild LRRo, with 35 speakers, 1.1k words, and spanning 21 hours, and Lab LRRo, with 19 speakers, 6.4k words, lasting about 5 hours. These datasets are divided into training, validation, and testing subsets.

3.1.4 Baseline systems

Lip-reading prediction in our study classifies each spoken word into a unique class. The classifier aims to either accurately identify the word's class or determine the top-k probable classes. We employ two primary models based on advanced deep neural network architectures: the VGG-M framework [10] and the Inception-V4 network [54], both well-regarded in classification tasks. VGG-M is implemented using the MT structure from [13].

Table 3.1 Performance results for the baseline models trained using the LRRo datasets.

Model architecture		MT				Inception-V4			
Dataset configuration		Test		Train-val		Test		Train-val	
	Accuracy	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Wild LRRo	21 classes	33%	61%	37%	68%	33%	62%	40%	64%
Lab LRRo	16 classes	76%	97%	80%	97%	75%	97%	80%	98%
	32 classes	81%	95%	82%	95%	77%	96%	81%	96%
	48 classes	71%	90%	71%	91%	71%	92%	71%	93%

Results

We trained our models using the train-val-test divisions. The highest performance for Romanian lip-reading models was observed with the Lab LRRo dataset. This outcome was somewhat anticipated since this dataset is meticulously managed. On the other hand, the Wild LRRo dataset, reflective of internet data, has considerable variations. Our best results show a top-1 accuracy of 33% for Wild LRRo and 71% for Lab LRRo. More details are provided in Table 3.1.

3.1.5 Conclusions

This section outlines our methodology, experiments and results for compiling and preparing the LRRo dataset for Romanian automatic lip-reading. LRRo, a word-level benchmark in a foreign language, includes data from uncontrolled environments sourced from YouTube and controlled lab settings for higher quality. We employed two notable architectures as baselines for our dataset: VGG-M, a Multiple Tower model and Inception-V4, adhering to the Inception framework.

3.2 Toward Language-independent Lip Reading: A Transfer Learning Approach

3.2.1 Introduction

While studies have extended lip-reading to non-English languages [18, 68, 26], their efficacy still lags behind English-focused systems. Our work leverages existing lip-reading knowledge, applying transfer learning to adapt insights from established domains to new languages. We introduce the LRM (Lip Reading Multilingual) dataset, a modest-sized collection integrating various pre-existing lip-reading datasets.

[3]the dataset is stored at the following link: https://doi.org/10.5281/zenodo. 3753559. For the lab recorded data, user permission was obtained and user data are anonymized. The wild data was retrieved from data that is already publicly available on the Internet.

3.2.2 Proposed Multilingual Learning Approach

Multilingual dataset generation

Table 3.2 Overview of the generated subsets.

Name	Split	Samples (hard)	Samples (hard+easy)	Pose	Env.
LRW	train	19,070	44,657	Natural	TV
	val	973	2,285		
	test	958	2,257		
LRRo	train	846	4,264	Natural/ Controlled	TV, Lab
	val	120	572		
	test	121	550		
LRW-1000	train	9,950	17,288	Natural	TV
	val	865	1,573		
	test	995	1,745		
LRM	train	-	66,209	Natural/ Controlled	TV, Lab
	val		4,430		
	test		4,552		

In our sequences, every image was processed using the MTCNN face detector [60] for facial landmark detection. To address variances in multilingual datasets, we standardized them by padding each to a uniform length.

Our evaluation used three word-level datasets in different languages: LRW (English), LRW-1000 (Mandarin), and LRRo (Romanian). Word selection was based on documented recognition challenges, classifying words as *easy* or *hard*. The composition of language subsets and the multilingual dataset is detailed in Table 3.2.

English subset

For the English language variant, we utilized the Lip Reading Dataset LRW [13]. The *easy* subset encompasses a total of 21,001 utterances, while the combined *easy* + *hard* subset has 49,199 utterances.

Mandarin subset

For the Mandarin language variant, we utilized the LRW-1000 dataset [64]. The *easy* subset encompasses 11,810 utterances, while the *easy* + *hard* subset totals 20,606 utterances.

Romanian subset

For the Romanian language variant, we employed the LRRo Dataset [26]. The *easy* subset comprises 1,087 utterances, while the combined *easy* + *hard* subset has 5,386 utterances.

Multilingual dataset

We have constructed a multilingual dataset by amalgamating data subsets from English, Mandarin, and Romanian. The resultant dataset, termed LRM, encompasses a lexicon of 141 words and a sum of 75,191 utterances.

Baseline

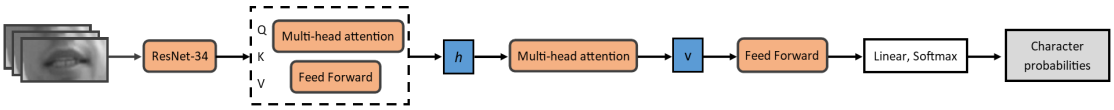


Fig. 3.3 Visual attention autoencoder model. The extracted lips sequences are processed by a spatio-temporal ResNet. On every decoder layer, the context vectors (h) are attended to separately by independent multi-head attention modules. The vectors are concatenated (v) and fed to feed-forward layers.

For our model construction, we followed the methodology outlined in [3], using a soft-attention autoencoder combined with ResNet-34 architecture. To address the challenges associated with multilingual datasets, we also utilized the improved D3D network, as described in [52], to tackle the complexities of multilingual datasets. This network represents an evolution of the model from [64], notably integrating 3D CNN layers at the front end instead of the initially suggested 2D CNN layers.

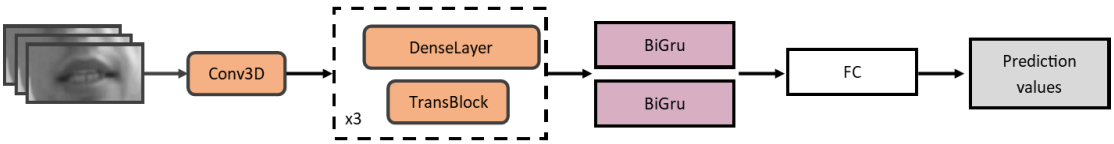


Fig. 3.4 D3D model. The visual cues are transformed by the 3D convolutional layers into spatio-temporal features. Feeding them through the gated recurrent units (GRUs), the final prediction values are generated.

3.2.3 Experimental setup

In the experimentation process, we leveraged the D3D network and the Autoencoder, employing the Adam optimization algorithm.

In the course of our experiments, we consistently evaluate performance using two primary metrics. The first of these is the recognition accuracy (Top-1 and Top-5). The second metric we employ is the Kappa coefficient, often denoted as κ.

Table 3.3 Classification accuracy on the LRM dataset. In bold are the best results.

Model	Invdividual subset	Acc. Top-1			
		hard		hard+easy	
		Val	Test	Val	Test
D3D	LRM	0.625	0.620	0.752	0.745
	LRW	0.835	0.830	0.836	0.847
	LRRo	0.578	**0.614**	0.885	0.892
	LRW-1000	0.463	0.418	0.535	**0.496**
Visual attention autoencoder	LRM	0.602	0.597	0.751	0.746
	LRW	0.843	0.841	0.842	0.844
	LRRo	0.513	0.510	0.898	**0.897**
	LRW-1000	0.452	0.440	0.514	0.498

3.2.4 Results

Baseline

When individually trained and tested on each dataset, both the D3D and Autoencoder models yielded the following Top-1 accuracy results for the hard configuration: 0.889 and 0.893 on the LRW subset, 0.594 and 0.373 on the LRW-1000 subset, and 0.342 and 0.333 on the LRRo subset, respectively. For the combined easy+hard configuration, their performances were 0.832 and 0.820 on the LRW subset, 0.373 and 0.393 on the LRW-1000 subset, and 0.418 and 0.629 on the LRRo subset, in the same order.

Transfer Learning

The English subset, with 1,000 utterances per class, is the most comprehensive. Therefore, it served as the main domain for fine-tuning to enhance the generalization capability of the evaluation models in relation to the other subsets.

Multilingual Learning

Our research, detailed in Table 3.3, demonstrated that a multilingual approach, using the LRM subset, excelled over transfer learning in the LRRo hard category and other subsets, highlighting the efficacy of mixed multilingual datasets.

Moreover, our evaluation, as shown in Figure 3.5, revealed that while words with 6 to 10 characters were the most common, our model maintained consistent accuracy across various word lengths, indicating its performance is not dependent on the word length.

3.2.5 Conclusions

This study addressed language-independent lip-reading dataset challenges using a multilingual subset, the D3D network, and a visual attention autoencoder, highlighting the effectiveness of these methods in cross-lingual speech recognition via transfer learning.

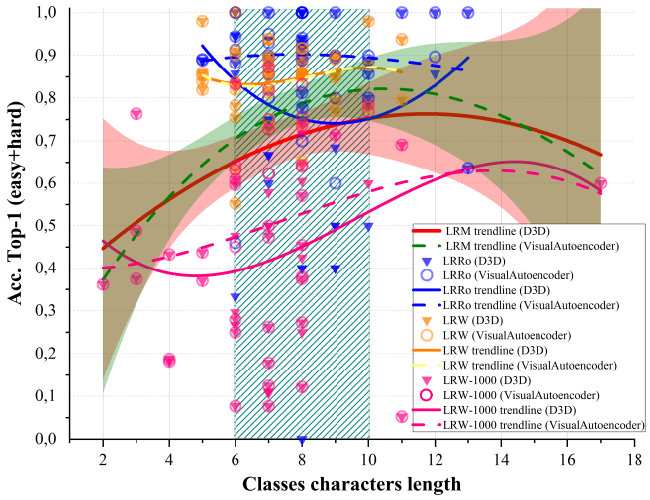


Fig. 3.5 Classification accuracy across word lengths. [30]

It also found that word length does not impact multilingual learning performance, indicating its wide linguistic applicability.

Chapter 4

Object Detection and Scene Understanding

Aerial imagery's object detection and scene comprehension are essential in today's interconnected world, offering unique insights across vast areas. The sheer scale and lower resolution of objects, along with diverse terrains, shadows, and lighting variations, complicate detection. Crowd analysis from the air is even more challenging but crucial for understanding migrations, event dynamics, emergencies, and urban movements. It's not just about counting but understanding collective behaviour in broader societal and environmental contexts. This research aims to tackle these technical challenges and delve into their wider implications.

4.1 Deep Learning-based Object Searching and Reporting for Aerial Surveillance Systems

Within the current computational landscape, the evolving dynamics of vehicle and aircraft fleets call for an integrated analytical approach. Geospatial imagery, covering both oblique and nadir perspectives, yields extensive data crucial for comprehensive area analysis.

4.1.1 Introduction

This work delves into the application of deep learning in aerial image analysis, relevant for environmental monitoring, infrastructure assessment, car impact analysis, airport traffic study, aircraft fleet management, and efficient automated interpretation of remote sensing images.

Our system for object search and reporting comprises two principal phases: (i) the detection of aircraft and (ii) the accumulation of statistical data.

Aircraft detection

Many recent deep learning studies on aerial and satellite imagery focus on different tasks. Etten's work [57] developed YOLT, a YOLO adaptation for large satellite images, using 416-pixel windowed segments. Zhou et al. [69] created a Multiscale Detection Network (MSDN) with a Deeper and Wider Module (DAWN) to better detect small aircraft and reduce background noise.

Statistical data output

In [42], training subset preparation is based on statistical data. [56] uses object statistical dimensions and runtime (km^2/min) to evaluate model efficacy. Essential object parameters, especially in emergency contexts, include aircraft dimensions, number, processed region, cluster size, and key-point positions.

Data collections for aircraft detection

A common aspect of the referenced studies is their use of specialized datasets, a necessity due to the scarcity of high-resolution satellite images with over 50 cm Ground Sample Distance (GSD). This lack has led to the creation of new datasets for specific tasks. Researchers have augmented training datasets from 50 cm GSD imagery, such as DOTA [61], Airbus Aircraft Detection [5], and PlanesNet [23], with artificially added airplanes to standard overhead images, examples being RarePlanes [50] and CGI Planes in Satellite Imagery [2].

4.1.2 Proposed object searching and reporting system

This section details our proposed model for aircraft detection and localization, designed for our recognition and reporting system and introduces a statistical method to derive key parameters from detected objects.

Object detection algorithms

SSD (Single Shot MultiBox Detector) network [40] performance, established the benchmark for our aircraft detection endeavor.

YOLOv4, as proposed by [8], has evidenced its prowess in terms of speed and accuracy on the MS COCO dataset [39].

The **YOLOv5** architecture, developed in PyTorch [31], maintains the backbone and network neck of YOLOv4. However, it integrates a distinct *Yolo layer* comprised solely of 1x1 convolutional layers.

Following this, our study explores the effectiveness of transfer learning in overcoming the hurdles of scarce, well-labeled data, crucial for developing a dedicated intra-class YOLOv5 model.

Data collection

Class *A* encompasses 443 objects, which include fighter jets, military transportation aircraft, and military helicopters, while class *B* has 231 objects, namely airliners, civilian transportation aircraft, and charter planes.

4.1.3 Experimental setup

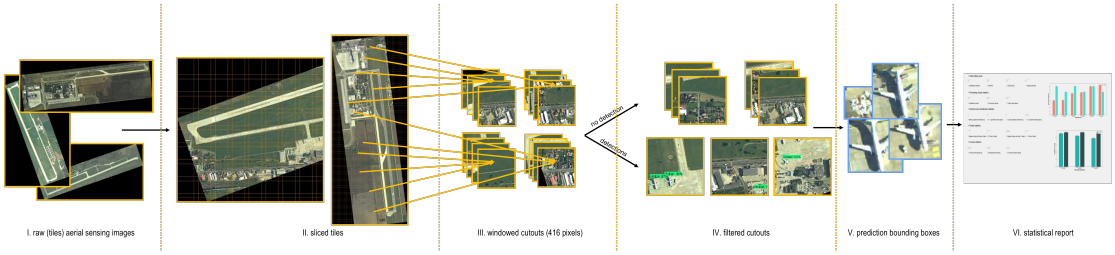


Fig. 4.1 Aircraft detection and reporting in aerial surveillance systems. [27]

Searching and reporting method

Figure 4.1 outlines the systematic process for an aircraft detection and reporting system, commencing from the initial aerial image tiles and 416-pixel chip creation, through to the extraction of prediction bounding boxes and the generation of statistical summaries.

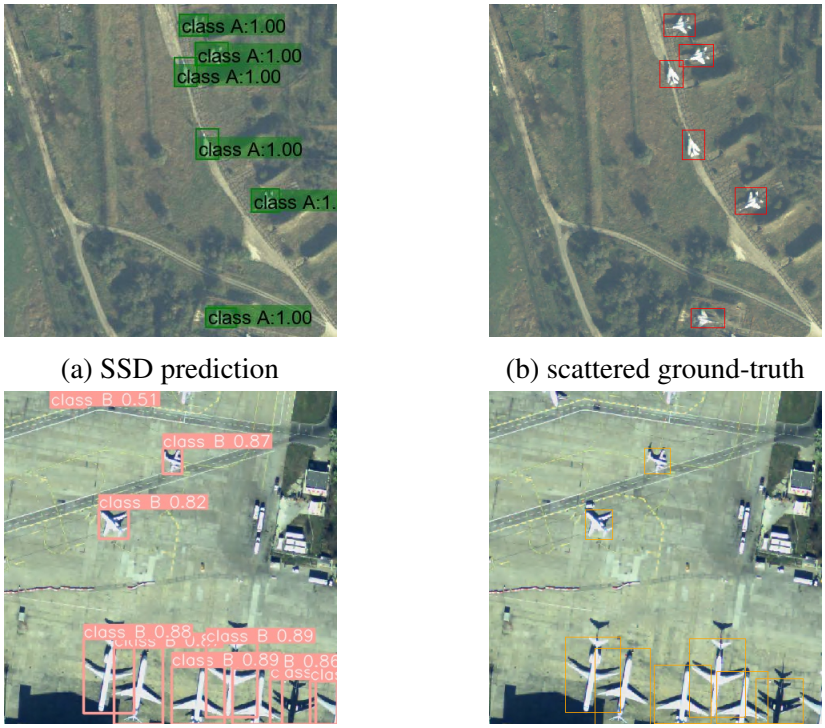
Evaluation protocols

We evaluate our models using mAP@50 (IoU and confidence thresholds at 0.5), F1-score and the confusion matrices.

4.1.4 Results

Baseline

Training occurred on the *train* subset and evaluation on the *test* subset, where the SSD network achieved mAP@50 scores of 0.690 for class A and 0.840 for class B, detailed in Table 4.1. These results provide a baseline for comparing enhancements by YOLO-based architectures.



(a) SSD prediction

(b) scattered ground-truth

(c) YOLOv5 prediction

(d) yin-yang pairs ground-truth

Fig. 4.2 Predictive outcomes for complex scenarios examined (the right column showcases 416 x 416 pixel cutouts; red bounding boxes indicate class A, while orange ones signify class B). [27]

Ablation experiments

Our experiments with various YOLOv5 configurations (*s, m, x*, and *l*) demonstrate their effectiveness in object detection and classification from remote sensing imagery, enhanced by transfer learning. It is vital to examine detection and localization performance

Table 4.1 Experimental results

Architecture	Pre-trained model	Batch size	No. of epochs	mAP@50		F1-score	Parameters
				class A	class B		
SSD	-	8	50	0.690	0.840	0.677	63.9M
YOLOv4	-	8	120	0.452	0.850	0.706	24.7M
YOLOv5s	-	8	200	0.792	0.886	0.800	7.2M
YOLOv5m	-	8	250	0.756	0.865	0.790	21.2M
YOLOv5x6l	YOLOv5l6	8	200	0.762	0.848	0.820	86.7M
YOLOv5x6l	YOLOv5l6	16	150	0.809	0.824	0.820	86.7M

across diverse aerial image scenarios, including characteristics such as chained, hidden, isolated, multi-class, varied scales, dispersed, tiny, and yin-yang pairs.

Table 4.1 shows models with parameters from 7.2M to 86.7M, where increased complexity and hyperparameters led to higher mAP@50 scores compared to a pre-trained model. These models were tested on unprocessed Google Earth images with 0.5 to 2.0 m GSD. While numerical accuracy is key, visual results, as displayed in Figure 4.2, are also crucial, showing predictions from SSD and YOLOv5 models on our specialized dataset. In summary, YOLOv4, being more compact, shows more false negatives, SSD underperforms in class *A* compared to YOLOv5x6l, but all models perform similarly in class *B*, with YOLO5s achieving the highest mAP@50 of 0.886.

The searching and reporting model

For our emergency use case, we fine-tuned YOLOv5x using the pre-trained YOLOv5l6 model, initially trained on the MS COCO dataset for 300 epochs. By applying transfer learning and training for 150 epochs with a batch size of 16, we attained our highest mAP@50.

4.1.5 Conclusions

We studied the YOLOv5x6, SSD, and YOLOv4 for aerial emergency analysis over large areas, using a custom two-class remote sensing dataset. Starting with SSD, then YOLOv4, and finally YOLOv5x6, we achieved mAP@50 scores of 0.809 for class A and 0.824 for class B, comparable to some one-class detectors. Our system processes aerial images in 416-pixel segments, handling a 50 km^2 area in 6 seconds with our best model. We evaluated the system using conventional metrics like mAP@50 and F1-score, as well as visual assessments in various scenarios, providing comprehensive insights for decision-making.

4.2 High Density Crowd Scene Detection in Untrimmed Streaming Videos for Surveillance Purpose

4.2.1 Introduction

Our study focuses on monitoring increasingly dense urban crowds, essential for security, through counting, detection, and segmentation. We enhance deep learning with basic image processing, considering crowd speed and direction, and categorize densities based on [19] and [6].

Crowd Counting Approaches

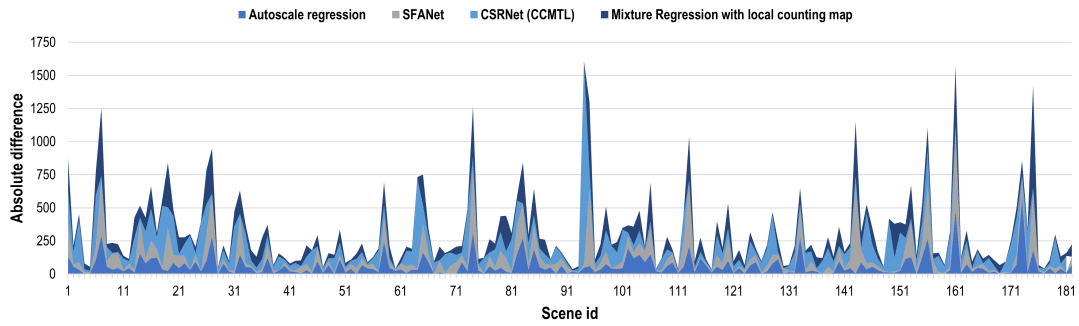


Fig. 4.3 Crowd counting techniques were employed as an initial filtering mechanism in the scene prediction methodology. The assessment was conducted on a test subset of the ShanghaiTech-A dataset, comprising 182 scenes. [29]

The main goal of visual crowd counting methods is to estimate the number of people in images or video frames. Our focus is on combining various counting strategies to average counts during frame processing, using the UrbanEvent data set, with the effectiveness of these methods demonstrated in Figure 4.3 based on ShanghaiTech-A scene evaluations.

Autoscale Regression. Xu et al. [62] introduces the Learning to Scale (L2S) module for pixel-wise crowd density mapping and a dynamic cross-entropy loss accounting for distance-label map geometry, improving accuracy on imbalanced datasets.

M-SFANet. In [55] is introduced the Modified-SFANet (M-SFANet), an advanced SFANet using VGG16-bn for enhanced multi-scale object and occlusion handling, employing CAN (context-aware) and ASPP (atrous spatial pyramid pooling) modules for precise crowd density mapping.

CSRNet. The CSRNet [38] employs VGG-16's first 10 layers for feature extraction and a dilated CNN for saliency, using bilinear interpolation and the final conv1-1-1 layer to produce the density map without pooling or deconvolutional layers.

Mixture Regression with Local Counting Map. The research by [41] presents a new training objective, the Local Counting Map (LCM). This objective integrates

a scale-aware module with a mixture regression and an adaptive soft interval module, improving counting regression on localized image sections.

Clustering for UrbanEvent Scene Categorization. To categorize diverse scenes from the UrbanEvent dataset, we used the K-Means clustering algorithm from *scikit-learn*. We employed feature embeddings from the pre-trained ResNet-152 model in the Pytorch framework, trained on the ImageNet1K dataset. Clusters ranged from 3 upwards, with lower counts grouping scenes by crowd density.

Overview of data collections for crowd analysis

ShanghaiTech-A. This component of the ShanghaiTech dataset consists of 1,198 images, primarily sourced from the internet.

WorldExpo. Also referred to as WorldExpo'10, this dataset was curated with the intent of augmenting the efficiency of crowd counting systems by offering a cross-scene dataset. It comprises 3,980 scenes sourced from the Shanghai 2010 World-Expo.

UCF-QNRF. The dataset addresses the shortcomings in crowd density of previous datasets, comprising 1,535 scenes.

JHU-CROWD++. A comprehensive, unconstrained crowd counting dataset, it includes 4,372 images.

4.2.2 Proposed approach

Our methodology detects dense crowds in raw streaming videos, consisting of (1) a crowd crounting technique for initial footage filtering, and (2) a scene variation detector to isolate scenes with unique viewer perspectives and crowd dynamics. See Figure 4.4 for a schematic representation.

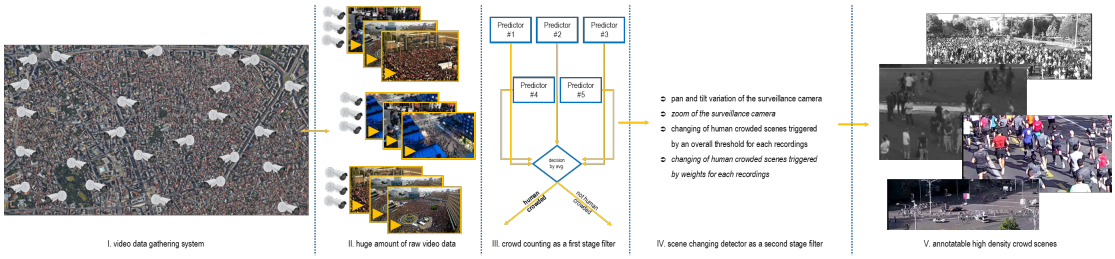


Fig. 4.4 Proposed methodology for human crowd scene analysis in urban contexts. [29]

Unprocessed data collection

We processed 47GB of real-world video footage, resulting in the UrbanEvent dataset with over 3.8 million frames, representing diverse urban aerial scenarios. The dataset includes an aggregate of 388,468 heads, with counts ranging from 2 to 1,370 heads per scene,

captured under various lighting, distances, occlusions, and crowd activities. However, UrbanEvent remains proprietary and isn't publicly available as of the publication date.

First stage - crowd counting task

We applied multiple crowd counting techniques, such as Autoscale Regression, M-SFANet, CSRNet, and Mixture Regression with Local Counting Map, on our untrimmed video dataset, training all models with the ShanghaiTech-A dataset. Based on averaged results, scenes were labelled as either *human crowded* or *not human crowded*. The subsequent analysis accounted for crowd dynamics and camera movements.

Second stage - proposed crowd scene change detector

Algorithm 1: Custom scene change detector for untrimmed video recordings

```
function custom_changeDetector (contentValue_set);
  Input  : CV - computed contentValue for given scenes
  Output : scene_outID - crowd unique scene indexes
  Init   : maxM = max { CV };
           dist_step init;
           ref_index init;
           δ init;
while CV not end do
    if crtCV > maxM − δ and crtCV_index ≥ ref_index + dist_step then
        add crtCV_index to the scene_outID list;
        ref_index = crtCV_index;
    end
    return scene_outID;
end
```

Due to the limitations of the initial K-means clustering based on feature embedding, we incorporated an additional filtering step using our custom crowd scene change detector, building upon the PySceneDetect FOOS's functionalities for video segmentation [9].

Our custom detector identifies variations in camera movement, zoom, and crowd changes in single frames, marked as (*scene_outID*), by comparing frame-to-frame HSV colour space metrics and pre-computed content values (*CV*) to set thresholds for crowd scene detection.

In the initial phase, we set a tolerance, δ, at 0.005, informed by content value analysis (Figure 4.5). Variations beyond a threshold, calculated as the local maxima (*maxM*) minus δ, are noted. The *dist_step* parameter helps reduce false negatives by tracking consecutive scenes for repeated changes, enabling efficient detection of zoom or camera movements without needing metadata.

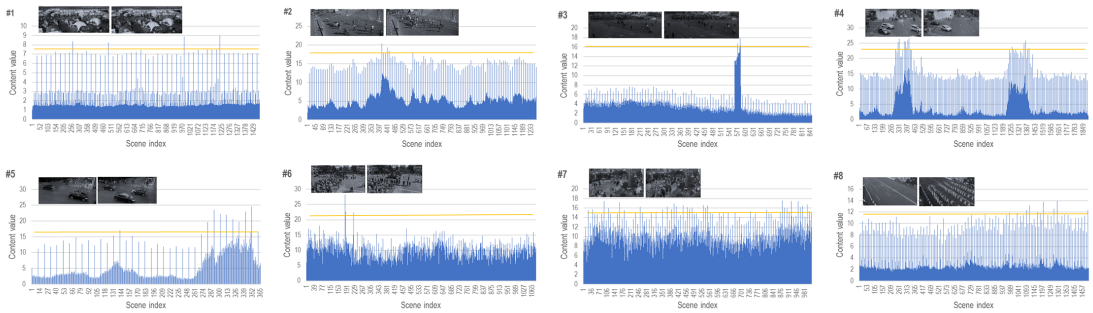


Fig. 4.5 The plots show the content values calculated by our detector for each scene pair in eight different event types: sporting events (#1, #2), mass gatherings (#3, #6), open street events (#4, #5, #7) and artistic events (#8). Peaks exceeding the threshold (orange line) signal scene extraction from the batch. The upper-left quadrant of each plot presents two adjacent scenes, illustrating the detector's decision-making. For example, in scenario #3, there's a pan-axis camera adjustment, while in #6, the crowd moves towards the camera. [29]

4.2.3 Experimental results

Our evaluation of the proposed crowd scene detection method includes both quantitative (using Mean Absolute Error - MAE and Mean Square Error - MSE for crowd counting models) and qualitative (focusing on identifying true-negative scenes) assessments. The method performs well across various video sequences, effectively averaging crowd counts even with minor occlusions and maintaining scale invariance in aerial camera views. M-SFANet showed the best performance in our tests with an MAE of 59.69, followed by Mixture Regression with LCM (MAE: 61.59), Autoscale Regression (MAE: 65.8), and CSRNet (MAE: 68.2). An expert-supervised refinement could further enhance dataset accuracy.

The outcomes of the K-means clustering algorithm and the original PySceneDetect (*content-aware* and *adaptive content detector* types) were similar, showing insensitivity to complex scenes categorized as crowded, not crowded or crowded with cars or other objects. In contrast, our proposed methodology effectively isolated distinct urban scenes with consistent crowd dynamics. Out of 3.8 million frames, it selectively identified only 2,031 crowded scenes, encompassing even blurry ones.

4.2.4 Conclusions

Our study developed a system combining four crowd counting methods and a scene change detector, successfully extracting 2,031 unique scenes from 3.8 million surveillance video frames. This approach, however, necessitates expert review to mitigate noise from dynamic elements, as our content analysis and algorithm suggest.

4.3 A Collection of Still Images for Coherent Crowd Analysis

4.3.1 Introduction

Advancements in computer vision methodologies have profoundly augmented public safety by identifying potential threats, unusual behaviours, forbidden items, and choke points in dense settings.

Our research highlights the value of analyzing sub-groups in crowds for sociobiological insights using segmentation methods. Current literature often neglects supervised crowd boundary delineation. We present a dataset of static images designed for consistent crowd segmentation to aid in understanding crowd dynamics [25]. Our approach is event-agnostic.

Crowd segmentation approaches

Wang et al. [59] and Zhang et al. [66] focus on crowd segmentation methods, leveraging spatial data and deep learning strategies to classify crowds based on various metrics. Mihalic [43] and Yang et al. [63] concentrate on pedestrian segmentation using motion-based attributes and spatial considerations in videos and static images. Abdullah et al. [1] and Khan et al. [33] explore crowd counting, behaviour analysis, and semantic segmentation techniques for distinguishing foreground and background, enhancing accuracy. For our project on predicting crowd boundaries, we utilize the U-Net architecture [46].

Crowd datasets

We've expanded on the work in [29], annotating the UrbanEvent dataset for semantic segmentation. Prominent static image datasets for crowd analysis include ShanghaiTech Part A and B [67], WorldExpo'10 [59], and JHU-CROWD++ [51].

4.3.2 Segmentation on still images

Proposed good practices and UrbanEvent data cleaning

Over 2,000 urban video feed frames were selected for annotation using a specialized scene change detection method [29]. The annotation process, involving the Computer Vision Annotation Tool (CVAT) by OpenCV [48], was conducted manually and validated by computer vision experts, excluding about 8% of frames from the final dataset.

In the delineation of crowd groupings, the following best practices were adhered to:

- (i) A pedestrian cluster is defined under specific conditions: either two isolated pedestrians or a group noticeably separate from a larger crowd.

- (ii) Occlusions are classified as background and excluded from annotations.
- (iii) Masks that display overfitting are not allowed.
- (iv) If spatial resolution prevents distinguishing individuals, assigning groups as background, especially if a pedestrian doesn't fit a 6x3 pixel box.

Benchmark algorithm for crowd segmentation

We used a Pytorch U-Net architecture for static crowd analysis, resizing UrbanEvent scenes to 1,280 x 1,920 pixels and dividing 1,884 annotated scenes in training and testing sets in an 8:2 ratio. The model was trained with pre-trained weights from the 2017 Carvana Image Masking Challenge, adapted for crowd segmentation using the UrbanEvent dataset.

4.3.3 Results on UrbanEvent benchmark

Our research assessed two main points: (i) the U-Net framework's accuracy in inferring crowd masks from open-source datasets and (ii) the performance measured by the Dice Similarity Coefficient (DSC). Initially trained on the Carvana dataset, the U-Net model reached a DSC of 0.9514 in just 2 epochs, which further trained on UrbanEvent subsets achieved a DSC of 0.5848 after 23 epochs. Examples from UrbanEvent, JHU-CROWD++, GCC, UCF-QNRF, and ShanghaiTech-A datasets are shown in Figure 4.6.

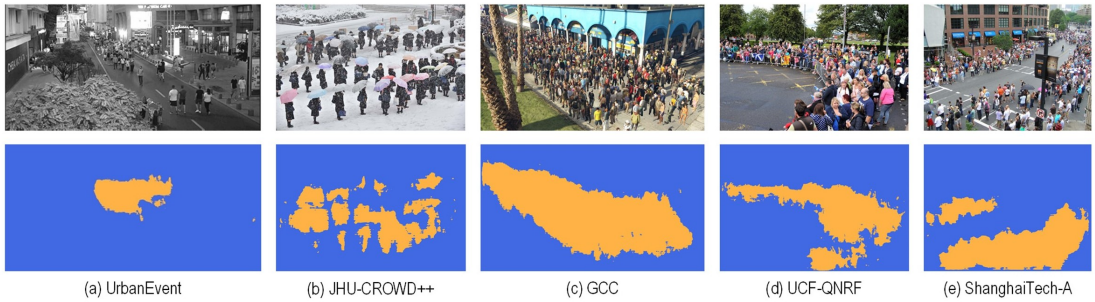


Fig. 4.6 Five sample pairs from crowd datasets and their predicted masks were chosen, including: (a) an UrbanEvent open-street event, (b) a JHU-CROWD++ image of humans in rows, (c) a synthetic GCC entry for crowd counting, (d) a UCF-QNRF public gathering frame, and (e) a densely populated scene from ShanghaiTech-A. [28]

4.3.4 Conclusions

We applied a supervised U-Net model to the UrbanEvent dataset, enhanced with adapted ShanghaiTech samples, achieving a 0.5848 Dice coefficient. Preliminary analysis suggests supervised learning with annotated crowds offers improved segmentation maps.

Chapter 5

Collaborative Findings on Human-to-media interaction analysis

In this study, we harness comprehensive data from the MediaEval Predicting Video Memorability task, aiming to pinpoint the characteristics that influence a video's memorability predictability.

On another front, the remarkable advancements in Generative Adversarial Networks (GANs) have ushered in high-definition, near-perfect image generation. Our approach, in this context, proposes a novel deepfake training strategy that embeds artificial markers within models.

5.1 Assessing the Difficulty of Predicting Media Memorability

5.1.1 Introduction

Memory research, spanning psychology, physiology [45], and fields like machine learning [49], focuses on image memorability [24] and its assessment [34]. The MediaEval Predicting Video Memorability task [53] (PVM) exemplifies the field's growth, attracting significant participation and setting benchmarks in memorability prediction.

Our research investigates what makes a video easy or hard to predict in computer vision, offering the following advancements over existing literature:

- We analyzed participants' runs from the 2022 MediaEval PVM task to identify challenging videos for automated prediction.

- We identify and assess various video attributes to provide a clear summary of the factors that affect the ease or difficulty of classifying a video's memorability.
- We explore how human-annotated ground truth data correlates with the effectiveness of different participant methods on the dataset.

5.1.2 Proposed approach

In the 2022 PVM task (subtask 1), participants submitted 33 runs. Our goal is to introduce a proficiency metric, detailing the systems, pre-processing steps, foundational principles of the metric and video classification methodologies.

The 2022 PVM dataset

The 2022 PMV task and its accompanying dataset [53] introduce a collection consisting of 10,000 three-second short videos. These videos, annotated for short-term memorability, are derived from the Memento10k dataset [44].

The official evaluation metric

Throughout its five editions, the MediaEval PVM task consistently employed the Spearman's rank correlation metric as the official measure for evaluating the efficacy of individual systems.

Video classes

We categorize the videos into two distinct groupings:

Quartile-Based Grouping: Videos are segmented into equal quartiles, labeled as $Q1$, $Q2$, $Q3$, and $Q4$. Each of these quartiles consists of 375 videos.

Threshold-Based Grouping: Videos are divided based on specific thresholds, resulting in segments labelled as $T1$, $T2$, $T3$, and $T4$. Specifically, the $T1$ group consists of videos where $0 \leq \hat{D}_i < 0.25$, $T2$ includes videos where $0.25 \leq \hat{D}_i < 0.5$, and so on. The respective sizes of these groupings are 787, 558, 130, and 25. Notably, the majority of videos fall within the $T1$ and $T2$ categories, indicating higher predictability.

Visual representations of these groupings are provided in Figure 5.1.

We propose using three feature sets to characterize videos and represent video categories: visual attributes, object attributes and correlations between human evaluations and prediction performance.

Fig. 5.1 Selected samples from each video category, representing the four quartiles and corresponding to four distinct thresholds. [14]

Proposed visual features

We analyze three frames per video using feature computation functions, averaging these for each video and then across each video category. Our process includes computing sharpness with the Laplacian [58] (f_1) and Canny operators (f_2), assessing colour vibrancy (f_3), contrast in RGB [32] (f_4) and average pixel values in the HSL colour model for hue (f_5), saturation (f_6), and lightness (f_7). We also include a video-level dynamism descriptor (f_8).

Proposed object-based features

Building on studies that connect object presence in images with subjective qualities such as interestingness [15], our hypothesis is that specific objects may positively or negatively influence prediction method effectiveness. We utilize a Mask R-CNN-based video annotation architecture [16] to track the frequency of the five most common objects and measure their coverage in the frame, creating two unique object-focused features (f_{10}).

Proposed annotator-based features

Our goal is to examine the link between video categories and their memorability values, based on human assessments. We analyze histograms of videos evenly spread across quartiles ($Q1$, $Q2$, $Q3$, and $Q4$) to observe category distributions and correlate them with inherent memorability using Spearman's coefficient.

5.1.3 Results

Employing the defined set of 11 features, we aim to clearly analyze and interpret the results they produce for the chosen video categories.

Visual features

The findings related to the eight visual features (f_1 - f_8) are delineated in Table 5.1.

Object-based features

The dataset's most common objects are "person", "chair", "car", "dining table", and "bird". Videos without identifiable objects are given a separate classification, detailed in Table 5.1.

Table 5.1 We analyzed the percentage differences in visual features (f_1 - f_8), the top five objects and object coverage feature (f_{10}) between the challenging quartiles ($Q2$ - $Q4$) and baseline quartile ($Q1$) and between threshold intervals ($T2$ - $T4$) versus the baseline interval ($T1$).

Feature	Q2	Q3	Q4	T2	T3	T4
f_1	30.58%	40.82%	30.12%	20.67%	18.11%	32.48%
f_2	16.67%	18.59%	11.36%	5.67%	4.99%	31.25%
f_3	-6.54%	-4.25%	-4.11%	-3.62%	-1.53%	3.76%
f_4	15.54%	17.44%	18.65%	9.35%	3.27%	5.01%
f_5	-1.05%	-0.21%	-0.39%	-0.35%	2.42%	10.97%
f_6	0.54%	6.51%	2.59%	3.32%	1.55%	5.14%
f_7	-6.84%	-6.21%	-6.43%	-3.55%	-0.66%	-1.14%
f_8	-1.67%	-10.91%	-10.01%	-5.51%	-18.38%	-38.13%
$f_9 - pers$	5.63%	2.63%	1.12%	-1.99%	-3.55%	9.11%
$f_9 - none$	27.69%	19.94%	47.97%	-3.77%	32.67%	-27.1%
$f_9 - chair$	-4.21%	-8.28%	-12.5%	5.73%	-29.67%	-
$f_9 - car$	-21.1%	-31.55%	-10.65%	-7.41%	-35.73%	-
$f_9 - table$	92.24%	199.46%	99.46%	25.36%	45.77%	94.4%
$f_9 - bird$	62.91%	138.02%	62.91%	92.14%	74.64%	-
f_{10}	-7.78%	-12.08%	-10.56%	-7.44%	-11.09%	-12.34%

Annotator-based features

Statistically, two trends emerge: (i) average memorability scores mostly come from $Q2$ or $Q3$ categories, and (ii) high or low memorability scores are predominantly in $Q1$ or $Q4$, with most unpredictable films in $Q1$ showing memorability indices between 0.7 and 0.9.

5.1.4 Conclusions

Our study explored how certain video features influence the difficulty of classifying videos by memorability scores, using data from the 2022 MediaEval PVM task. We found that harder-to-classify videos often had higher contrast and sharpness but lower dynamism, less brightness and vibrancy but more saturation, fewer major objects and medium memorability scores, suggesting that focusing on these traits during training could improve classifier effectiveness for challenging samples.

5.2 Deepfake Sentry: Harnessing Ensemble Intelligence for Resilient Detection and Generalisation

5.2.1 Introduction

Our study investigates how various perturbations affect deepfake detection efficiency, examining three datasets: FaceForensics++ (FF++) [47], DeepFake Detection Challenge Preview Dataset (DFDC Preview) [17] and Celeb-DF [37]. We also introduce an autoencoder ensemble approach to mitigate these perturbation effects.

The key contributions are:

- Introduction of two universal augmentations to enhance compromised detectors.
- Demonstration of our method's effectiveness through extensive testing.
- Detailed evaluation of deepfake detectors against various perturbations.

The goal of deepfake detection is to differentiate real media from synthetic ones. There are two main approaches: high-level and low-level detection, both being actively developed for this purpose.

1. High-level deepfake detection

High-level forensic methods focus on meaningful attributes like behavioural anomalies, such as missing physiological cues in synthesized videos. For instance, [35] addresses this issue and [36] exploits a limitation in DeepFake videos related to face image size constraints.

2. Low-level deepfake detection

Low-level forensic methods detect pixel-level inconsistencies, with [65] identifying unique fingerprints in GAN training and [22] highlighting the importance of temporal coherence in certain spatial regions.

Our paper unveils the potential vulnerabilities of forensic classifiers when subjected to a variety of adversarial tactics that compromise their prediction accuracy.

5.2.2 Proposed approach

In the next section, we introduce our neural ensemble approach for deepfake detection, outlined in Figure 5.2. Our training process involves facial region extraction, perturbations, artificial fingerprint generation and embedding, enhancing model robustness. The modified faces are then analyzed to detect deepfake, making our method easily adaptable across various models.

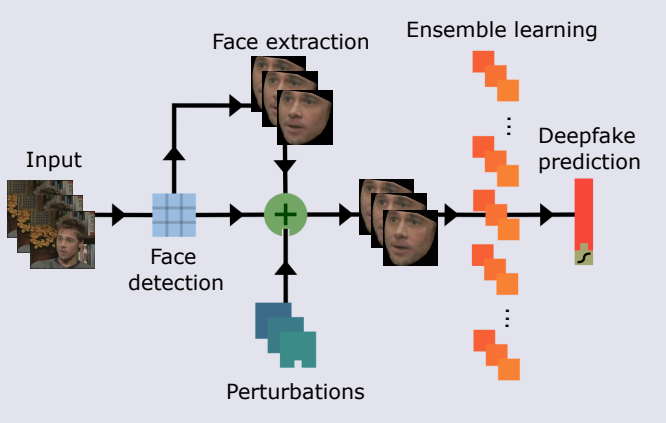


Fig. 5.2 The figure outlines our deepfake detection training framework with four stages: (1) facial region extraction, (2) perturbation application, (3) artificial fingerprint integration using autoencoders, and (4) predictive modelling for deepfake detection. [70]

Neural ensemble architecture for deepfakes

Our goal is to develop a resilient deepfake detection strategy that can withstand advancements in generative models by integrating artificial fingerprints into our models. We employ a set of image-forensic autoencoders, each transforming an image and passing it through an ensemble of classifiers to produce a scalar real-valued output. A higher output value indicates a higher likelihood of the input image being a synthetic or fabricated one.

Attacks against DeepFake classifiers

Our study explores how deepfakes affect image-forensic classifiers and their resistance to adversarial manipulation. We introduce an adversarial function that deceives the classifier, leading to potentially inaccurate authenticity judgments. We analyze the classifier's resilience through: (1) white-box attacks, granting full parameter transparency, and (2) black-box attacks, with no such insights, as well as real-world-inspired image modifications, such as basic perturbations and JPEG compression.

5.2.3 Results

Experimental setup

Our research primarily focuses on assessing the practical impact of the autoencoder-based augmentation algorithm. To achieve this, our experiments were designed with three key objectives: (i) using simple, state-of-the-art models that emphasize autoencoder-based augmentation over specific model architectures; (ii) ensuring our models are versatile and perform consistently across various deepfake techniques; (iii) training autoencoders to closely replicate original dataset images, resulting in nearly imperceptible variations.

Our approach takes a frame-level perspective, fine-tuning the XceptionNet network [11] on the FF++ dataset for binary classification of genuine and deepfake images. We evaluate its universality on the CelebDF and DFDC preview datasets, using $3 \times 299 \times 299$ facial images. We employ 80 autoencoder models, including convolutional autoencoders and U-Net architectures, as described earlier. Images are processed through autoencoders and undergo various augmentations, including sharpness, contrast, perspective changes, affine transformations, blurring, rotations, color variations, and noise. Additionally, we apply JPEG lossy compression at different levels (10, 20, 30, 50, and 80).

Datasets

FF++ is a prominent benchmark for deepfake detection, offering an unprocessed collection of 700 videos.

The **DFDC Preview** includes 4,113 DeepFake videos, featuring 66 individuals from diverse demographic backgrounds.

Celeb-DF dataset contains 590 authentic and 5,639 DeepFake videos, totalling over two million frames.

Experimental results

In our analysis, we measure detection effectiveness by calculating the area under the receiver operating characteristic curve (AUC) for 16 consecutive frames and then average the frame-level scores to make a video-level decision.

In our experiments, we evaluate the effectiveness of four deepfake classifier configurations: (i) **Baseline Model (BL)**, (ii) **Classic Augmentation Model (CA)**, (iii) **Ensemble Augmentation Model (EA)** and (iv) **Ensemble Augmentation + Classical Augmentation Model (EA+CA)**. These models are fine-tuned on the FF++ dataset and tested for generalizability on CelebDF and DFDC preview datasets. Table 5.2 demonstrates how these models perform when exposed to various perturbations across the datasets, assessing their adaptability and resilience to practical image processing scenarios.

Table 5.3 enumerates the AUC outcomes of deepfake detection across varying JPEG compression intensities.

DeepFake attacks. As seen in prior studies, our results confirm the vulnerability of forensic classifiers to these adversarial attacks. White-box attacks significantly reduce the AUC from around 0.9x to below 0.1x, sharply contrasting the AUC of 0.5, indicating chance-level classification. When access to classifier parameters is restricted in black-box attacks, the ROC decreases by at least 5% across all datasets. In black-box attack scenarios, our algorithm's resilience is linked to the performance of the employed model, with the conventional ResNet-50 model used for attack generation.

5.2.4 Conclusions

Our study's primary focus was on utilizing an ensemble autoencoder approach to augment data for deepfake detection. We used basic yet cutting-edge models to highlight the significance of our augmentation method, emphasizing adaptability and versatility in the training process. These autoencoders were trained to capture subtle deep learning patterns that may not be apparent to human observers. Our evaluation aimed to ascertain whether our augmentation technique improves adaptability, resilience to data alterations, tolerance to compression, and resistance to adversarial attacks. The outcomes reveal

Table 5.2 In the evaluation of deepfake detection, the AUC scores were determined under various perturbations.

Perturbation	**FF++**				**CelebDF (generalization)**				**DFDC (generalization)**			
	BL	CA	EA	EA+CA	BL	CA	EA	EA+CA	BL	CA	EA	EA+CA
No distortion	0.995	0.996	0.997	**0.999**	0.748	0.770	0.805	**0.826**	0.697	0.709	0.715	**0.722**
Adjust Sharpness	0.969	0.991	0.993	**0.995**	0.728	0.731	0.767	**0.769**	0.691	0.699	**0.713**	**0.713**
Autocontrast	0.968	0.990	0.993	**0.994**	0.650	0.654	0.691	**0.701**	0.698	0.701	0.703	**0.714**
Random Perspective	0.930	0.963	0.972	**0.989**	0.691	0.697	0.720	**0.722**	0.670	0.670	**0.709**	**0.709**
Color Jitter	0.961	0.986	0.989	**0.994**	0.741	0.764	0.796	**0.799**	0.670	0.680	0.699	**0.708**
Random Resized Crop	0.860	0.967	0.976	**0.978**	0.681	0.647	0.692	**0.713**	0.646	0.628	0.688	**0.701**
Gaussian Blur	0.962	0.987	0.990	**0.998**	0.765	0.778	**0.826**	0.824	0.682	0.676	**0.716**	0.713
Random Noise	0.975	0.984	0.991	**0.995**	0.731	0.764	0.803	**0.823**	0.661	0.682	0.692	**0.704**
Random Rotation	0.956	0.983	0.985	**0.992**	0.745	0.765	0.783	**0.785**	0.665	0.671	0.698	**0.710**
Random Affine (A)	0.810	0.844	0.860	**0.906**	0.612	0.620	0.658	**0.698**	0.652	0.631	0.656	**0.676**
Random Affine (B)	0.915	0.949	0.959	**0.966**	0.697	**0.731**	0.716	0.726	0.688	0.689	0.673	**0.699**
Average	0.936	0.967	0.973	**0.982**	0.708	0.720	0.750	**0.762**	0.674	0.676	0.696	**0.706**

Table 5.3 Performance outcomes for deepfake detection, measured in terms of AUC, were assessed under distinct JPEG compression levels.

JPEG Compression	**FF++**				**CelebDF (generalization)**				**DFDC (generalization)**			
	BL	CA	EA	EA+CA	BL	CA	EA	EA+CA	BL	CA	EA	EA+CA
10	0.700	0.861	0.839	**0.866**	0.583	0.521	0.530	**0.574**	**0.675**	0.662	0.661	0.661
20	0.765	0.930	0.931	**0.933**	0.586	0.634	0.632	**0.640**	0.666	0.673	**0.699**	0.697
30	0.783	**0.965**	0.962	**0.965**	0.553	0.671	0.668	**0.677**	0.652	0.697	0.682	**0.701**
50	0.839	0.981	0.980	**0.983**	0.590	0.690	0.711	**0.716**	0.645	0.685	0.694	**0.706**
80	0.943	0.990	0.990	**0.992**	0.710	0.763	0.764	**0.779**	0.665	0.682	0.702	**0.719**
Average	0.806	0.945	0.940	**0.947**	0.604	0.655	0.661	**0.677**	0.660	0.679	0.687	**0.697**

Table 5.4 Results for deepfake detection, quantified in AUC, under both blackbox and whitebox attack scenarios.

Adversarial Attacks	**FF++**				**CelebDF (generalization)**				**DFDC preview (generalization)**			
	BL	CA	EA	EA+CA	BL	CA	EA	EA+CA	BL	CA	EA	EA+CA
Black-box	0.887	0.898	0.911	**0.926**	0.645	0.712	0.681	**0.717**	0.597	0.635	0.643	**0.660**
White-box	**0.014**	0.003	0.009	0.011	0.001	0.014	0.027	**0.029**	0.010	0.073	0.069	**0.078**

a significant enhancement in deepfake detection effectiveness in practical scenarios through our strategy.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In my doctoral research, I have contributed significantly to computer vision, real-world data development, and deep learning-based processing pipelines.

Part I of my thesis covers the basics of speech recognition, aerial monitoring, crowd analysis, and AI in human-to-media interaction.

Part II details my contributions in these areas, offering insights for computer vision engineers to replicate and apply these methods in various scenarios. Key chapters include Chapter 3 on visual speech recognition and lip reading, Chapter 4 on object detection and scene understanding, and Chapter 5 on human-to-media interaction, focusing on memorability and deepfake detection.

6.2 Original contributions

- [C1] introduces a new dataset for coherent crowd analysis using deep learning, derived from the UrbanEvent unlabeled data collection. It aims to improve crowd segmentation methodologies and provides guidelines for creating trainable datasets. The study evaluates the performance of AI methods in different urban environments and highlights their effectiveness in accurately defining crowd boundaries, particularly in previously unseen scenes, thereby establishing a baseline for this task.
- In [C2] numerous runs from the 2022 MediaEval Predicting Video Memorability task, are exploited to identify challenging movies for automated predictions and investigate video features affecting memorability. It evaluated prediction algorithms against human-annotated data, finding that harder-to-classify videos had higher contrast, sharpness, and saturation, but lower brightness, colour vibrancy, and object coverage. Addressing these factors in classifier training could enhance performance on difficult data.
- [C3] tackles the efficient management of large-scale surveillance video data, focusing on extracting unique pedestrian cluster instances. It introduces a two-phase methodology: firstly, implementing various crowd counting methods using regression techniques, assessing their effectiveness in urban settings; secondly, employing a new scene change detection method based on HSV colour space differences, inspired by PySceneDetect's scene transition identification techniques. From 3.8 million sampled frames of challenging videos, this approach successfully extracted 2,031 crowd scenes for the UrbanEvent data collection. This study aims to refine and direct future research in practical surveillance applications.
- [J1] introduces an ensemble-based autoencoder for data augmentation, in the context of advancements by Generative Adversarial Networks in creating high-quality images. The goal was to train models with wide adaptability for effective deepfake detection in various real-world situations. Tested on benchmark datasets for deepfake detection, this combined approach of the ensemble autoencoder and traditional augmentation techniques consistently surpassed other strategies in Area-under-the-curve scores, irrespective of the perturbations, compression rates, and adversarial methods (both white-box and black-box) evaluated.
- [C4] outlines a methodology for aircraft detection in aerial surveillance, offering statistical analysis and visualizations for two categories of identified aircraft. It highlights spatial interactions between objects and employs a deep learning architecture effective in complex scenarios. The study focuses on optimizing object detection for systems with limited GPU resources, crucial for analyzing large terrains in high-traffic, time-sensitive situations. The system provides decision-makers

with data on aircraft dimensions, quantity, surveyed area, clustering intensity, and critical coordinates, calibrated for high performance with common Google Earth imagery resolutions.

- [C5] introduces a methodology to improve lip reading systems, combining data augmentation and curriculum learning. It focuses on developing a language-independent system, especially for languages with limited research resources. The approach involves a multilingual learning technique and the creation of the Lip Reading Multilingual dataset (LRM), featuring Romanian, English, and Mandarin. Words in the dataset are classified as 'easy' or 'hard' based on recognition challenges. Applying this method to two deep learning models showed significant performance enhancements, indicating the potential for cross-lingual knowledge transfer.
- [C6, R1] introduces a new method for visual speech recognition (VSR) in under-resourced languages, focusing on Romanian. It features the first Romanian VSR dataset, the Lip Reading Romanian dataset (LRRo), compiled from controlled and natural settings. The LRRo dataset, available online since 2020, contains recordings from over 50 speakers and annotations for more than 7,000 words from various sources. The study details the methodology for collecting and analyzing lip reading data, including an examination of Romanian visemes. It also reports on the development and testing of two deep learning models for word-level speech recognition, marking significant progress in creating the first Romanian visual speech recognition system.

6.3 Perspectives for further developments

This thesis explored human interactions through AI, focusing on real-world applications and the challenges in integrating AI into daily tasks. It emphasized the need for research on case-insensitive algorithms, particularly in surveillance and security. The findings show a gap between conventional AI metrics and practical needs, suggesting improvements in deep learning architectures. A key issue identified was the difficulty in managing large datasets during training, highlighting the need for efficient learning methods to develop more resource-efficient models without starting from scratch for each new scenario.

6.4 List of original publications

Journal

J1: L.-D. Ştefan, C. Stanciu, M. Dogariu, M.G. Constantin, **A.-C. Jitaru**, B. Ionescu: Deepfake Sentry: Harnessing Ensemble Intelligence for Resilient Detection and

Generalisation. In University Politehnica of Bucharest Scientific Bulletin Series C-Electrical Engineering and Computer Science, 72(1):11-27, 2023. [70]

Conference Papers

C1: **A.-C. Jitaru**, B. Ionescu: A collection of Still Images for Coherent Crowd Analysis, In Proceedings of 1st Doctoral Symposium on Electronics, Telecommunications & Information Technology, CEUR Workshop Proceedings, CEUR-WS, 2023. [28]

C2: M.G. Constantin, M. Dogariu, **A.-C. Jitaru**, B. Ionescu: Assessing the Difficulty of Predicting Media Memorability. 20th International Conference on Content-based Multimedia Indexing - CBMI, September 20-22, Orleans, France, 2023. [14]

C3: **A.-C. Jitaru**, B. Ionescu: High Density Crowd Scene Detection in Untrimmed Streaming Videos for Surveillance Purpose, 2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 1-6, 2023. [29]

C4: **A.-C. Jitaru**, C.-E. Barbu, B. Ionescu: Deep Learning-based Object Searching and Reporting for Aerial Surveillance Systems, 2022 14th International Conference on Communications (COMM), Bucharest, Romania, 2022, pp. 1-7, doi: 10.1109/COMM54429.2022.9817266. [27]

C5: **A.-C. Jitaru**, L.-D. Ştefan, B. Ionescu: Toward Language-independent Lip Reading: A Transfer Learning Approach, IEEE International Symposium on Signals, Circuits and Systems – ISSCS, ISBN: 978-1-6654-4942-7,
DOI: 10.1109/ISSCS52333.2021.9497405, July 15-16, Iași, Romania, 2021. [30]

C6: **A.-C. Jitaru**, Ş. Abdulamit, B. Ionescu: LRRo: a Lip Reading Data Set for The Under-Resourced Romanian Language, In Proceedings of the 11th ACM Multimedia Systems Conference (MMSys '20), Association for Computing Machinery, New York, NY, USA, 267–272. https://doi.org/10.1145/3339825.3394932. [26]

Research Project

R1: 2018–2020: researcher, project SPIA-VA. "Technologies and Innovative Video Systems for Person Re-Identification and Analysis of Dissimulated Behavior", owner Polytechnic University of Bucharest, partners UTI Grup, Romanian Ministry of National Defence—Military Equipment and Technologies Research Agency, public beneficiary Protection and Guard Service, funded by UEFISCDI, Solutions Axis, grant 2SOL/2017 (budget 2.2M Eur).

References

[1] Abdullah, F. and Jalal, A. (2023). Semantic segmentation based crowd tracking and anomaly detection via neuro-fuzzy classifier in smart surveillance system. *Arabian Journal for Science and Engineering*, 48(2):2173–2190.

[2] aceofspades914 (2019). Cgi planes in satellite imagery w/ bboxes.

[3] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., and Zisserman, A. (2018a). Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727.

[4] Afouras, T., Chung, J. S., and Zisserman, A. (2018b). Lrs3-ted: a large-scale dataset for visual speech recognition. *ArXiv*, abs/1809.00496.

[5] Airbus DS GEO, S. A. (2021). Airbus aircraft detection. sample aircraft detection dataset from airbus high resolution satellite imagery.

[6] Aldayri, A. and Albattah, W. (2022). Taxonomy of anomaly detection techniques in crowd scenes. *Sensors*, 22(16).

[7] Assael, Y. M., Shillingford, B., Whiteson, S., and de Freitas, N. (2016). Lipnet: Sentence-level lipreading. *ArXiv*, abs/1611.01599.

[8] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.

[9] Castellano, B. (2022). *PySceneDetect Documentation*.

[10] Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. *ArXiv*, abs/1405.3531.

[11] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

[12] Chung, J. S., Senior, A. W., Vinyals, O., and Zisserman, A. (2016). Lip reading sentences in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453.

[13] Chung, J. S. and Zisserman, A. (2016). Lip reading in the wild. In *ACCV*.

[14] Constantin, M. G., Dogariu, M., Jitaru, A.-C., and Ionescu, B. (2023). Assessing the difficulty of predicting media memorability. In *Proceedings of the 20th International Conference on Content-based Multimedia Indexing*. IEEE.

[15] Dhar, S., Ordonez, V., and Berg, T. L. (2011). High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011*, pages 1657–1664. IEEE.

[16] Dogariu, M., Stefan, L.-D., Constantin, M. G., and Ionescu, B. (2020). Human-object interaction: Application to abandoned luggage detection in video surveillance scenarios. In *2020 13th International Conference on Communications (COMM)*, pages 157–160. IEEE.

[17] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., and Ferrer, C. C. (2019). The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*.

[18] Faisal, M. and Manzoor, S. (2018). Deep learning for lip reading using audio-visual information for urdu language. *CoRR*, abs/1802.05521.

[19] Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., and Zhu, C. (2015). Fast crowd density estimation with convolutional neural networks. *Engineering Applications of Artificial Intelligence*, 43:81–88.

[20] Georgescu, A.-L., Cucu, H., and Burileanu, C. (2017). Speed's dnn approach to romanian speech recognition. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–8. IEEE.

[21] Georgescu, A.-L., Cucu, H., and Burileanu, C. (2019). Kaldi-based dnn architectures for speech recognition in romanian. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1–6. IEEE.

[22] Guan, J., Zhou, H., Hong, Z., Ding, E., Wang, J., Quan, C., and Zhao, Y. (2022). Delving into sequential patches for deepfake detection. In *Advances in Neural Information Processing Systems*, volume 35, pages 4517–4530. Curran Associates, Inc.

[23] Hammell, B. (2018). Planes in satellite imagery. detect aircraft in planet satellite image chips.

[24] Isola, P., Parikh, D., Torralba, A., and Oliva, A. (2011). Understanding the intrinsic memorability of images. *Advances in neural information processing systems*, 24.

[25] Japar, N., Kok, V. J., and Chan, C. S. (2021). Coherent group detection in still image. *Multimedia Tools and Applications*, 80:22007–22026.

[26] Jitaru, A. C., Abdulamit, Ş., and Ionescu, B. (2020). Lrro: a lip reading data set for the under-resourced romanian language. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 267–272.

[27] Jitaru, A.-C., Barbu, C.-E., and Ionescu, B. (2022). Deep learning-based object searching and reporting for aerial surveillance systems. In *2022 14th International Conference on Communications (COMM)*, pages 1–7. IEEE.

[28] Jitaru, A.-C. and Ionescu, B. (2023a). A collection of still images for coherent crowd analysis. In *Doctoral Symposium on Electronics, Telecommunications Information Technology*. CEUR Workshop Proceedings, CEUR-WS.

[29] Jitaru, A.-C. and Ionescu, B. (2023b). High density crowd scene detection in untrimmed streaming videos for surveillance purpose. In *2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6. IEEE.

[30] Jitaru, A.-C., Ştefan, L.-D., and Ionescu, B. (2021). Toward language-independent lip reading: A transfer learning approach. In *2021 International Symposium on Signals, Circuits and Systems (ISSCS)*, pages 1–4. IEEE.

[31] Jocher, G., Stoken, A., Borovec, J., Chaurasia, A., Changyu, L., Laughing, V., Hogan, A., Hajek, J., Diaconu, L., Kwon, Y., et al. (2021). ultralytics/yolov5: v5. 0-yolov5-p6 1280 models aws supervise. ly and youtube integrations. *Zenodo*, 11.

[32] Ke, Y., Tang, X., and Jing, F. (2006). The design of high-level features for photo quality assessment. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 419–426. IEEE.

[33] Khan, K., Khan, R. U., Albattah, W., Nayab, D., Qamar, A. M., Habib, S., and Islam, M. (2021). Crowd counting using end-to-end semantic image segmentation. *Electronics*, 10(11):1293.

[34] Khosla, A., Raju, A. S., Torralba, A., and Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, pages 2390–2398.

[35] Li, Y., Chang, M.-C., and Lyu, S. (2018a). In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE.

[36] Li, Y. and Lyu, S. (2019). Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[37] Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216.

[38] Li, Y., Zhang, X., and Chen, D. (2018b). Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1091–1100.

[39] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing.

[40] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer.

[41] Liu, X., Yang, J., and Ding, W. (2020). Adaptive mixture regression network with local counting map for crowd counting. *CoRR*, abs/2005.05776.

[42] Mao, J., Zhang, X., Ji, Y., Zhang, Z., and Guo, Z. (2021). Improved high precision aircraft target detection method of yolt. In *Journal of Physics: Conference Series*, volume 1955, page 012028. IOP Publishing.

[43] Mihalić, D., Marčetić, D., and Ribarić, S. (2020). An approach to crowd segmentation at macroscopic level. In *2020 International Symposium ELMAR*, pages 105–108. IEEE.

[44] Newman, A., Fosco, C., Casser, V., Lee, A., McNamara, B., and Oliva, A. (2020). Multimodal memorability: Modeling effects of semantics and decay on video memorability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 223–240. Springer.

[45] Phillips, W. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics*, 16:283–290.

[46] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

[47] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

[48] Sekachev, B., Manovich, N., Zhiltsov, M., Zhavoronkov, A., Kalinin, D., Hoff, B., TOsmanov, Kruchinin, D., Zankevich, A., DmitriySidnev, Markelov, M., Johannes222, Chenuet, M., a andre, telenachos, Melnikov, A., Kim, J., Ilouz, L., Glazov, N., Priya4607, Tehrani, R., Jeong, S., Skubriev, V., Yonekura, S., vugia truong, zliang7, lizhming, and Truong, T. (2023). opencv/cvat: v2.5.0.

[49] Shekhar, S., Singal, D., Singh, H., Kedia, M., and Shetty, A. (2017). Show and recall: Learning what makes videos memorable. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2730–2739.

[50] Shermeyer, J., Hossler, T., Etten, A. V., Hogan, D., Lewis, R., and Kim, D. (2020). Rareplanes: Synthetic data takes flight. *CoRR*, abs/2006.02963.

[51] Sindagi, V. A., Yasarla, R., and Patel, V. M. (2020). Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *Technical Report*.

[52] Stafylakis, T. and Tzimiropoulos, G. (2017). Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105*.

[53] Sweeney, L., Constantin, M. G., Demarty, C.-H., Fosco, C., de Herrera, A. G. S., Halder, S., Healy, G., Ionescu, B., Matran-Fernandez, A., Smeaton, A. F., and Sultana, M. (2023). Overview of the mediaeval 2022 predicting video memorability task. In *Working Notes Proceedings of the MediaEval 2022 Workshop*.

[54] Szegedy, C., Ioffe, S., and Vanhoucke, V. (2016). Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*.

[55] Thanasutives, P., Fukui, K.-i., Numao, M., and Kijsirikul, B. (2020). Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. *arXiv preprint arXiv:2003.05586*.

[56] Van Etten, A. (2018). You only look twice: Rapid multi-scale object detection in satellite imagery. *arXiv preprint arXiv:1805.09512*.

[57] Van Etten, A. (2019). Satellite imagery multiscale rapid detection with windowed networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 735–743.

[58] Wan, J., He, X., and Shi, P. (2007). An iris image quality assessment method based on laplacian of gaussian operation. In *MVA*, pages 248–251.

[59] Wang, Z., Cheng, C., and Wang, X. (2018). A fast crowd segmentation method. In *2018 International Conference on Audio, Language and Image Processing (ICALIP)*, pages 242–245.