UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE AUTOMAÇÃO E SISTEMAS

Flávio Gabriel Oliveira Barbosa

Detecting and Retrieving Actions in Still Images

Florianópolis
2023

Flávio Gabriel Oliveira Barbosa

Detecting and Retrieving Actions in Still Images

Tese submetida ao Programa de Pós-Graduação em Engenharia de Automação e Sistemas da Universidade Federal de Santa Catarina para a obtenção do título de Doutor em Engenharia de Automação e Sistemas.
Orientador: Prof. Marcelo Ricardo Stemmer, Dr. - Ing.

Florianópolis
2023

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Barbosa, Flávio Gabriel Oliveira
Detecting and Retrieving Actions in Still Images / Flávio Gabriel Oliveira Barbosa ; orientador, Marcelo Ricardo Stemmer, 2023.
125 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia de Automação e Sistemas, Florianópolis, 2023.

Inclui referências.

1. Engenharia de Automação e Sistemas. 2. Recuperação de Imagens Baseada em Conteúdo Semântico. 3. Recuperação de Imagens Baseada em Ações.. 4. Aprendizado Profundo.. 5. Redes Neurais Convolucionais.. I. Stemmer, Marcelo Ricardo . II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia de Automação e Sistemas. III. Título.

Flávio Gabriel Oliveira Barbosa

Detecting and Retrieving Actions in Still Images

O presente trabalho em nível de Doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Leandro dos Santos Coelho, Dr.
Pontifícia Universidade Católica do Paraná (relator)

Prof. Maurício Edgar Stivanello, Dr.
Instituto Federal de Santa Catarina

Prof. Eric Aislan Antonelo, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Doutor em Engenharia de Automação e Sistemas.

ICP-Edu
Documento assinado digitalmente
Julio Elias Normey Rico
Data: 10/08/2023 11:09:11-0300
CPF: ***.840.859-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

ICP-Edu
Documento assinado digitalmente
Marcelo Ricardo Stemmer
Data: 10/08/2023 15:00:21-0300
CPF: ***.093.279-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Marcelo Ricardo Stemmer, Dr. -Ing.
Orientador

Florianópolis, 2023.

Este trabalho é dedicado a Deus e à minha família

AGRADECIMENTOS

Gostaria de expressar minha imensa gratidão a todos aqueles que estiveram ao meu lado e me apoiaram durante essa jornada incrível.

Em primeiro lugar, agradeço a Deus por ser minha fonte de força e suporte em todos os momentos da minha vida. Sua presença constante iluminou meu caminho e me deu a coragem e a determinação para seguir em frente, mesmo nos momentos mais desafiadores.

Agradeço à minha família. À minha mãe, Nilmar, meu pai, Jorge, e minha irmã, Luciana, agradeço do fundo do meu coração. Vocês são a base de tudo o que conquistei até agora. Seu amor incondicional, apoio incansável e crença em mim foram os pilares que sustentaram minhas realizações. Sou verdadeiramente abençoado por ter vocês ao meu lado.

Não posso deixar de expressar minha gratidão ao meu avô Nilson. Seu apoio constante e orgulho evidente em relação à minha jornada acadêmica são uma grande motivação. Seu sorriso radiante após minha defesa foi uma das maiores recompensas que poderia receber. Obrigado por sempre acreditar em mim.

Aos meus amigos, meu sincero agradecimento. Durante o processo de escrever este texto, parei para refletir, listei treze pessoas e percebi o quão afortunado sou por amigos incríveis que desempenharam papéis fundamentais em minha jornada. Vocês estiveram ao meu lado, me incentivaram, compartilharam momentos de alegria e também me apoiaram nos momentos difíceis. Sou grato por cada risada, palavra de encorajamento e por serem pessoas verdadeiramente excepcionais em minha vida.

Por fim, gostaria de estender meus agradecimentos a todos os meus professores, com um agradecimento especial ao meu orientador, o Prof. Dr. -Ing. Marcelo Ricardo Stemmer. Sua orientação, conhecimento e apoio foram essenciais para minha formação acadêmica. Sua dedicação em compartilhar seu conhecimento e sua paixão pela área contribuíram significativamente para o meu crescimento como estudante e como pessoa. Serei eternamente grato por tudo o que aprendi com você.

A todos vocês, meu profundo agradecimento. Suas contribuições, apoio e presença em minha vida foram inestimáveis. Serei eternamente grato por ter cada um de vocês ao meu lado. Obrigado do fundo do meu coração.

“Non est ad astra mollis e terris via”
- Sêneca

RESUMO

O rápido crescimento do mercado de dispositivos móveis aliado ao crescente uso de redes sociais resultou em um aumento significativo no volume de imagens geradas e consumidas. Consequentemente, a busca por imagens em um grande banco de dados torna-se uma necessidade cada vez maior. Apesar de todos os avanços dos últimos anos, quando deseja-se buscar por imagens utilizando conceitos semânticos (recuperação semântica) ainda é um desafio na visão computacional. Os humanos são capazes de observar semelhanças abstratas e complexas em imagens. Na recuperação semântica, o objetivo da pesquisa raramente pode ser determinado com base unicamente na própria imagem da consulta, e traduzir esse conhecimento em processamento digital não é uma tarefa trivial. Esta tese propõe dois *frameworks* de recuperação de imagens baseados em ações para imagens estáticas. O primeiro, Act-CBIR, é um *framework* de duas fases composto por uma fase de Criação de Dicionário e outra fase de Recuperação de Imagem. Essas etapas são compostas por um módulo de Detecção de ações para detectar regiões de interesse (ROIs) e extrair características de cada respectiva ROI; um módulo de codificação e indexação de ações para representar cada ROI de maneira concisa; um banco de dados de índice para armazenar todas as ROIs em uma forma indexada; e um modelo de similaridade para recuperar imagens dadas uma imagem de consulta. Dado o aumento exponencial na quantidade de imagens sendo gerados o que traduz-se em bancos de dados cada vez maiores, propomos uma alternativa de codificação capaz de binarizar os códigos para usar a eficiente distância de hamming. Essa abordagem é comparada com dois outros pipelines de indexação de codificação: utilizar os códigos diretamente da camada totalmente conectada introduzida para esse fim e utilizar distância euclidiana, e também *Local Sensitive Hashing* (LSH) para recuperar imagens. Finalmente, o modelo de similaridade busca imagens por meio de uma classificação indireta usando o algoritmo *Quicksort*. Apesar de suas vantagens, esse primeiro *framework* não considera nenhuma informação adicional além da região de interesse de ação, tornando difícil até mesmo para nós, humanos, descrever algumas ações. A segunda abordagem, Act-Retrieval, é baseada em múltiplas entradas, detecção de ação, aprendizado por dicas e um módulo de atenção para superar esse problema. Para validar experimentalmente os dois conceitos, uma análise quantitativa é realizada utilizando as métricas de *mean Average Precision* (mAP) e AP@10, que leva em consideração somente as dez primeiras imagens retornadas, uma vez que muitas vezes estamos interessados apenas nos primeiros resultados de nossa consulta. Uma análise qualitativa também foi realizada, observando os mapas de características gerados e o resultado correspondente de cada estratégia para melhor ilustrar as diferentes informações absorvidas pelo segundo *framework*. Comparamos nossos resultados com trabalhos de referência e no estado-da-arte na área de recuperação de imagens, superando-os por larga margem. Portanto, esta tese contribui para reduzir a lacuna semântica considerando imagens estáticas e ações de recuperação de imagens.

Palavras-chave: Recuperação de Imagens Baseada em Conteúdo Semântico. Recuperação de Imagens Baseada em Ações. Aprendizado Profundo. Redes Neurais Convolucionais.

RESUMO ESTENDIDO

Introdução

O avanço das tecnologias portáteis causou um aumento exponencial no nosso contato com imagens digitais. Diariamente, visualizamos incontáveis imagens através de nossos celulares e computadores. Consequentemente, a busca por imagens estáticas em um grande banco de dados que corresponda a uma consulta tem ganhado destaque dada sua demanda cada vez maior. A maioria das imagens não possui metadados (como tags ou pistas contextuais), então os motores de busca de imagens que quantificam o conteúdo de uma imagem, como o *Content-Based Image Retrieval* (CBIR), são ideais para muitas aplicações. CBIR pode ser dividido em subcampos de acordo com o tema da pesquisa: (i) A recuperação da classe visa encontrar imagens da mesma classe da consulta. (ii) A recuperação de instância é mais desafiadora, pois atende a objetivos específicos para muitas aplicações. Trata-se da tarefa de encontrar imagens contendo o mesmo objeto que podem ser capturadas em diferentes condições, exigindo a correspondência de padrões visuais de granulação fina entre as imagens. (iii) A terceira categoria é a recuperação semântica e cobre um espectro mais amplo. A categoria neste tipo de pesquisa não significa necessariamente um objeto ou classe. O objetivo exato da pesquisa do usuário raramente pode ser determinado com base na própria imagem da consulta. Apesar de todos esses avanços, a recuperação de imagens baseada em conteúdo semântico ainda é um desafio. Os humanos procuram por semelhanças semânticas, mas o banco de dados só pode fornecer imagens semelhantes por processamento digital. Pesquisadores apontam uma lacuna entre a abundância de informações de alto nível que um ser humano pode extrair de uma imagem e a representação que as máquinas extraem, chamando esse fenômeno de "lacuna semântica". Humanos procuram semelhanças semânticas, mas o banco de dados só pode fornecer imagens semelhantes por processamento digital. Nesse sentido, pode-se citar as ações humanas, como “andar de bicicleta” ou “jogar tênis”, que fornecem uma descrição natural para muitas imagens estáticas. Além disso, a lacuna semântica entre as imagens e propriedades dos objetos limita amplamente a eficiência de recuperação, pois há inconsistência na compreensão de dados visuais para diferentes usuários, tornando desafiador eliminar essa lacuna semântica. Nesse contexto, essa tese aborda o problema de recuperação de imagens que podem ser descritas por ações humanas.

Objetivos

Essa tese tem como objetivo reduzir a lacuna semântica em recuperação de imagens por conteúdo, focando-se na recuperação de imagens que possuam pessoas realizando a mesma ação, sem nenhum auxílio de vídeos. Para esse fim, um dos objetivos da tese é realizar revisões sistemáticas da literatura, bem como realizar a recuperação de imagens sem caixas delimitadoras feitas manualmente em tempo de teste, propondo sistemas de recuperação de imagem baseado em ação sem explorar qualquer informação adicional proporcionada por humanos em tempo de teste. Além disso, dado o aumento exponencial do número de imagens sendo geradas a cada dia, estudar e introduzir um método de codificação eficaz (em termos de velocidade e custo

computacional) para recuperação de imagens. Por fim, outro objetivo importante dessa tese é atingir resultados superiores no tema que ela se propõe quando comparado a trabalhos de referência e no estado-da-arte em recuperação de imagens baseado em conteúdo.

Metodologia

A metodologia dessa tese inicia-se por revisões sistemáticas nos temas abordados. Foram realizadas revisões em reconhecimento de ações em imagens estáticas e em recuperação de imagens baseadas em ação que utilizam apenas imagens estáticas. Assim, foi possível identificar o estado atual da pesquisa nas áreas supracitadas, além de lacunas no estado-da-arte. O próximo passo consistiu em propor um sistema capaz de recuperar imagens baseando-se nas ações executadas através apenas de imagens estáticas, de forma eficiente em termos de precisão, mas sem o uso de caixas delimitadoras em tempo de teste, como a maioria dos trabalhos no estado-da-arte em reconhecimento de ações em imagens estáticas, o que limitaria o uso prático do sistema. Também foi introduzida uma estratégia para binarizar o vetor de características (FC-Bin) com a intenção de usar a distância de Hamming como uma medida de comparação de similaridade. Essa estratégia provou possuir uma relação de troca eficiente em velocidade e precisão, especialmente ao extrapolar conjuntos de dados maiores. Foi demonstrada a eficácia dessa abordagem tanto na precisão quanto no tempo de pesquisa. Apesar de todas as vantagens desse primeiro *framework*, foi possível notar que seu ótimo desempenho era relativamente menor para classes mais abstratas, nas quais se exige maior informação de contexto. De forma a melhorar ainda mais o sistema, introduzimos o segundo *framework*, que leva em consideração múltiplas informações, bem como algumas técnicas mais refinadas de pós e pré-processamento, a fim de obter melhor resultados.

Resultados

Avaliamos nossa abordagem nos conjuntos de dados públicos de reconhecimento de ação *People Playing Musical Instrument* (PPMI) e PASCAL VOC 2012, indicando um desempenho eficiente em termos de precisão para ambos os sistemas de recuperação de imagens baseados em ações utilizando imagens estáticas, principalmente quando compara-se o desempenho dos sistemas propostos com trabalhos de referência e no estado-da-arte em recuperação de imagens por conteúdo. A alternativa de binarização dos códigos para uso da eficiente distância de hamming para calcular similaridade de imagens também provou-se uma alternativa válida. Para melhor compreender as diferentes informações absorvidas pelo segundo *framework*, também foi realizada uma análise qualitativa, observando os mapas de características gerados e o resultado correspondente de cada estratégia. As análises mostraram, também, que o uso de múltiplas informações trouxe grande benefício principalmente na recuperação de imagens para ações mais abstratas, nas quais o contexto tem papel fundamental para seu entendimento. Assim, os resultados evidenciam que a presente tese trabalha no sentido de reduzir a lacuna semântica supracitada.

Considerações finais

Nessa tese, foram apresentados dois sistemas para recuperação de imagens por conteúdo baseados em ações e utilizando somente imagens estáticas. Ambas as abordagens são baseadas em detectar a ação através de um detector de objetos baseado em aprendizado profundo treinado para esse fim, e não utilizam caixas delimitadoras em tempo de teste, como a maioria dos trabalhos em reconhecimento de ações em imagens estáticas. A análise do primeiro *framework*, Act-CBIR, demonstra que usar a imagem inteira (cena) ou um detector de pessoa genérico não é ótima para a tarefa quando compara-se com o detector de ação, que captura interações humano-objeto. Também foi introduzida uma estratégia para binarizar o vetor de características com a intenção de usar a eficiente distância de Hamming para comparar a similaridade entre imagens, isto é, ações. Essa estratégia provou entregar um custo benefício eficiente em velocidade e precisão, especialmente ao extrapolar para conjuntos de dados maiores. Apesar das vantagens e resultados interessantes da abordagem anterior, o Act-CBIR desconsidera completamente qualquer informação adicional além da região de interesse da ação, tornando difícil até mesmo para nós, humanos, entendermos algumas ações nessa pequena região de interesse. O segundo sistema, Act-Retrieval, é uma estrutura para recuperação de imagens baseada em múltiplas entradas, a própria detecção de ação, aprendizagem por dicas e um módulo de atenção com o objetivo de inserir mais informação no sistema a fim de auxiliá-lo a encontrar imagens nas quais precisa-se de um grau de abstração maior, inserindo informações de contexto. Essa estrutura pode considerar várias informações como entrada e saída, melhorando o desempenho de recuperação em termos de precisão, especialmente para classes fortemente dependentes de informações de contexto, como "andar". Também foram introduzidas algumas melhorias através de uma abordagens de *hint learning*, combinando características semânticas globais e locais para obter uma melhor representação da ação. Assim, nossas abordagens propostas trabalham para reduzir a lacuna semântica considerando imagens estáticas e ações, tendo resultado superiores a trabalhos de referência e no estado-da-arte na área de recuperação de imagens, superando-os por uma larga margem.

Palavras-chave: Recuperação de Imagens Baseada em Conteúdo Semântico. Recuperação de Imagens Baseada em Ações. Aprendizado Profundo. Redes Neurais Convolucionais.

ABSTRACT

The rapid growth of the mobile device market combined with social media resulted in a significant increase in the volume of images being generated and consumed. Consequently, searching for still images in a large database that matches a query becomes an increasing necessity. Despite all the advances in the last years, semantic image retrieval is still a challenge in computer vision. Humans are capable of observing complex abstract similarities given single or multiple images. In semantic retrieval, the search objective can rarely be determined based on the query image by itself, and translating this knowledge into digital processing is not a trivial task. This thesis proposes two action-based CBIR frameworks that only consider still images. The first framework, Act-CBIR, is a two-staged framework composed of a Dictionary Creation stage and another stage of Image Retrieval. These stages are composed of an Action Detection module to detect regions of interest (ROIs) and extract features from each respective ROI; an Action Encoding and Indexing module to represent each ROI concisely; an Index database to store all ROIs in an indexed form, and a Similarity Model to retrieve images given a query image. Given the exponential increase in the size of the databases, this thesis proposes an encoding alternative able to binarize the codes to use the efficient hamming distance and compare with two other encoding indexing pipelines: computing codes directly from our introduced fully-connected feature layer and using cosine distance, and Locality Sensitive Hashing (LSH) to retrieve images. Finally, the similarity model retrieves results using an indirect sort using the Quicksort algorithm. Despite its advantages, the framework does not consider any additional information beyond the region of interest of action, making it difficult even for us humans to describe some actions. The second framework, Act-Retrieval, is based on multiple inputs, action detection, hint-learning, and an attention module to overcome this issue. To experimentally validate both concepts, a quantitative analysis is performed using the standard mean Average Precision (mAP), and the AP@10, since we are often interested only in the first results of our query. A qualitative analysis was also performed, observing the feature maps generated and the corresponding result of each strategy to better illustrate the different information absorbed by the second framework. The results are compared with reference and state-of-the-art works in the image retrieval field, surpassing them by a large margin. Therefore, this thesis contributes to reduce the semantic gap considering static images and actions for image retrieval.

Keywords: Semantic Content-Based Image Retrieval. Action-Based Image Retrieval. Deep Learning. Convolutional Neural Networks.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

ANN	Artificial Neural Network
AutoML	Automated Machine Learning
bbox	Bounding boxes
BoF	Bag of freebies
BoS	Bag of specials
BOW	Bag of Words
CASP	Critical Appraisal Skills Program
CBIR	Content-Based Image Retrieval
CBVR	Content-Based Video Retrieval
CHORs	Circular Histogram of Oriented Rectangles
CNN	Convolutional Neural Network
CONV	Convolutional
DBN	Deep Belief Network
DELWO	Deep ensemble learning based on the weight optimization
DL	Deep Learning
DPM	Deformable Part Model Detector
DPSH	Deep pairwise-supervised hashing
DSIFT	Dense SIFT
FC	Fully Connected
FPS	Frames per second
FV	Fisher Vectors
h-o	human-object
HOG	Histogram of Oriented Gradients
HRNet	High-Resolution Network
HUE	Robust Hue descriptor
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
KNN	K-Nearest Neighbors
LLC	Locality-constrained linear coding
LSH	Locality Sensitive Hashing
mAP	Mean Average Precision
MIT	Massachusetts Institute of Technology
MKL	Multiple Kernel Learning
MNIST	Modified National Institute of Standards and Technology
NAS	Neural Architecture Search
PCA	Principal Component Analysis
POOL	Pooling
PPMI	People Playing Musical Instrument

ReLU	Rectified Linear Unit
ROI	Region of Interest
ROIs	Regions of Interest
RPN	Region Proposal Network
RQ	Research question
SC	Shape Context
SIAR	Still Image Action Recognition
SIFT	Scale-invariant Feature Transform
SLR	Systematic Literature Review
SPM	Spatial Pyramid Matching
SR	Systematic Review
SSD	Single Shot MultiBox Detector
SSDH	Semi-supervised deep hashing
STIP	Spatial Point of Interest
SURF	Massachusetts Institute of Technology
SVHN	The Street View House Numbers
SVM	Support-Vector Machines
VDR	Visual Dependency Representation
VII	Visual Inverted Index
VLAD	Vector of Linearly Aggregated Descriptors
YOLO	You-only-Look-Once

CONTENTS

1 INTRODUCTION

This section serves as a compass, diligently guiding readers through the foundational aspects of this thesis. Firstly, the motivation that drove the initiation of this research is presented. Subsequently, the objectives of this thesis are defined, outlining the intended aims and desired outcomes. Additionally, the contributions made by this research to the existing body of knowledge are presented, highlighting their unique value, originality, and potential impact. Lastly, a comprehensive overview of the thesis structure is provided, offering readers a coherent roadmap of the document's organization and emphasizing the interconnectedness of its various sections.

1.1 MOTIVATION

In the early days of artificial intelligence, it was believed that solving the computer vision problem would be relatively straightforward. This is illustrated by "The summer vision project" announced by Seymour Papert and formed by a group of students at Massachusetts Institute of Technology (MIT) in 1966 (SZELISKI, 2010). The goal of the project was to build a significant part of the visual system to map the camera input to a description in terms of objects and background. This artificial vision system could be used as an input for high-level cognitive tasks such as reasoning and planning. Ever since, vision and image understanding have been studied from different perspectives. For example, from the field of cognitive psychology, aiming at understanding human perception, from the field of physics, focusing on modeling physical properties of light emission and surface reflections in images, and from the field of computer science, striving for automatic systems for various vision tasks, like 3D reconstruction and object recognition. The binding factor of these research interests is the use of computers to model and test different theories and methods. Decades later, we can only conclude that a complete understanding of the human visual system is still elusive, although we know some of its principles.

The difficulty is partly explained by the fact that scene understanding and object recognition are complex problems in which we try to recover an understanding of the world given single or multiple images. It is surprising that humans do this effortlessly with very high accuracy, despite the task's complexity. The visual world in its full complexity is challenging to model because of the enormous amount of different concepts, the infinite possibilities to project a 3-dimensional scene onto a 2-dimensional image plane, and due to complex scenes with different levels of occlusions and different lighting conditions. Furthermore, there is high intrinsic variability in the appearances of objects of the same class. This is even strengthened by visual ambiguity, when two different concepts have a similar appearance, and also by semantical ambiguity when a single concept has multiple meanings.

This thesis considers a fundamental problem of computer vision: enabling computers to see the way we see things. In the future, we wish our machines would match the capabilities of human vision. It is interesting to note that every second we receive a tremendous amount of visual data, and almost unconsciously, we process this information very quickly. Classifying an object as a table, a ball, or a scene as a mountain or a river is pretty trivial for us.

The advancement of portable technologies has inducted an exponential increase in images being generated and consumed daily. Consequently, the research on image retrieval has seen a considerable increase recently (JUN et al., 2019). Image search engines can be divided into three categories: meta-data search, content-based search, and a hybrid approach of the previous two. Meta-data search is similar to keyword-based search engines. Usually, it does not examine the image's content, relying on textual cues, such as tags and contextual cues that appear near the image. This additional information may not always be available, which weakens this approach. On the other hand, most images do not have meta-data (as tags or textual cues). Hence, image search engines that quantify the content of an image, as the Content-Based Image Retrieval (CBIR), become a necessity. CBIR has been an active field of study in recent years due to its enormous potential for applications, such as remote sensing, medical image search, object re-identification, shopping recommendation in online markets, and many others (ZHOU, W. et al., 2017; CHEN et al., 2021). Considering the success in applying Convolutional Neural Network (CNN) based methods in the image domain, many researchers have proposed to use them in the CBIR context.

Content-Based Image Retrieval can be divided into sub-fields according to the search theme: (i) Class retrieval aims to find images of the same class as the query. (ii) Instance retrieval is more challenging, as it satisfies specific objectives for many applications. It is the task of finding images containing the same object that may be captured under different conditions as the query image (CHEN et al., 2021), requiring matching fine-grained visual patterns between images. (iii) The third category is semantic retrieval, and it covers a more broad spectrum. The category in this type of search does not necessarily mean an object or class. The exact search objective of the user can rarely be determined based on the query image by itself (BARZ; DENZLER, 2021).

Despite all these advances, semantic content-based image retrieval is still a challenge. Barz and Denzler (2021) identify a gap between the high-level plethora of information that a human can extract from an image and the representation that machines extract. This gap between the richness of high-level human perception and low-level machine descriptions is known as the "semantic gap". Humans search for semantic similarity, but the database can only provide similar images by digital processing (ALZU'BI et al., 2015). In addition, the semantic gap between images and object properties broadly limits the retrieval efficiency as there is inconsistency in understanding

visual data for different users, making it challenging to eliminate (YUE et al., 2011).

Human actions provide a natural description of several static images. Human action recognition based on video has been a relatively well-studied research problem in computer vision when compared to still image action recognition. Unlike video-based action recognition, where temporal sequences are available and play critical roles, the area of still image-based action recognition requires image content interpretation (ZHANG et al., 2016). Despite the efforts made in recent years, the task of image-based action recognition remains a challenging problem due to factors as disordered backgrounds, occlusion, different viewpoints, variations in human posture, changes in lightning, and lack of movements (LIU, C. et al., 2014). Lately, the image-based approach has obtained increasing attention in the research community by creating datasets and challenges, as the PASCAL VOC Action Recognition (EVERINGHAM, Mark et al., 2015).

The most common approach to still image action recognition assumes that bounding boxes to map people's location are provided at both training and test time, which may discourage practical applications. The work of Sharma et al. (2012) employed the bag-of-words framework for action classification. As the human pose often plays a fundamental role in action recognition, another interesting approach is to find solutions for human pose estimation (THURAU; HLAVÁC, 2008). However, this approach is limited by the fact that similar poses can be associated with different actions, as a person brushing his teeth and other blowing bubbles can have similar body poses (YAN et al., 2017a). In contrast, others address the problem by taking a human-centric approach that localizes persons and then finds objects and their relationship to the person instance, as (YAO et al., 2011), that proposed a method that learns a set of sparse attributes and parts of people for action recognition, defining action attributes as the verbs that describe the properties of human actions, while the parts of actions are objects and poselets that are closely related to the actions. (KHAN, F. S. et al., 2018) proposed a method to scale coding within a bag of deep features framework, using absolute and relative scale information to encode in final image representations, leading to significant gains in action recognition performance. Due to the discriminative power of modern CNNs, Mohammadi et al. (2019) propose an ensemble learning approach using eight state-of-the-art CNNs. Safaei and Foroosh (2019) created an image representation that captures the spatial-temporal combined features and classifies actions accordingly to introduce motion information in still images. Nevertheless, these works concentrate mainly on the recognition task, disregarding the localization task, and real-world images where multiple human actions may be presented simultaneously, action detection plays a significant role, especially in practical applications such as CBIR.

Few studies address the problem of retrieving images based on actions. Piji Li et al. (2011) proposed a method based on Multiple Kernel Learning (MKL) to detect latent

action in an image using *hot-regions* with a sliding-window concept. (JONES; SHAO, 2013) presented a few approaches to action-based image retrieval using still image information, or combining data from video or text sources, using different methods, as Support-Vector Machines (SVM), Asymmetric Bagging and Random Subspace SVM (ABRS-SVMs), Simple Maximum of Similarities, and also representation methods, as vocabulary guided pyramid match, spatio-temporal pyramid match, and the original bag of words. The work of Elliott et al. (2014) models the spatial relationships between image regions using a structured image representation to distinguish between object co-occurrence and interaction, representing an image as a directed acyclic graph over a set of labeled object region annotations, thus, identifying latent representation of the depicted action in the image, achieving the best result in Mean Average Precision (mAP) slightly above 50% for the 2011 PASCAL VOC action detection dataset. Although these works tackle action-based image retrieval, none of these works study different encoding strategies nor explore promising deep learning approaches, which have been boosting results over several computer vision tasks.

This thesis proposes two action-based CBIR frameworks for still images. The first is a two-staged Act-CBIR method, and it demonstrates that generic person detectors or using the full image are suboptimal for the task of retrieving images based on actions. This framework is composed of a Dictionary Creation stage and another stage of Image Retrieval. These stages are composed of an Action Detection module to detect Regions of Interest (ROIs) and extract features from each respective Region of Interest (ROI); an Action Encoding and Indexing module to represent each ROI concisely; an Index database to store all ROIs in an indexed form, and a Similarity Model to retrieve images given a query image. Given the exponential increase in the databases, this thesis proposes an encoding alternative able to binarize the codes to use the efficient hamming distance and compare with two other encoding indexing pipelines: computing codes directly from the introduced fully-connected feature layer and using cosine distance, and Locality Sensitive Hashing (LSH) to retrieve images. Finally, the similarity model retrieves results using an indirect sort using the Quicksort algorithm. Despite its advantages, the framework does not consider any additional information beyond the region of interest of action, making it difficult even for us humans to describe some actions. The second framework, Act-Retrieval, is based on multiple inputs, action detection, hint-learning, and an attention module to overcome this issue. To experimentally validate both concepts, a quantitative analysis is performed using the standard mean Average Precision (mAP), and the AP@10, since we are often interested only in the first results of the query. A qualitative analysis was also performed, observing the feature maps generated and the corresponding result of each strategy to better illustrate the different information absorbed by the second framework.

1.2 THESIS OBJECTIVES

Regarding the motivation of this work and the current state-of-the-art of semantic image retrieval, especially retrieving images based on actions in still images, this thesis's main research objective is ***to propose an action-based image retrieval system using only still images*** and it is based on the following research questions (RQs):

RQ 1. Is it feasible to build an action-based image retrieval mechanism that considers only still images?
RQ 2. How to make this system viable for practical applications?
RQ 3. How to reduce the computational cost of the search due to the large dimension of the CNN feature representations?

This general objective carries within five specific sub-objectives:

(i) To perform systematic literature reviews about Still Image Action Recognition (SIAR) and action-based CBIR mechanisms to this end, listing relevant works from the past 5 years in this thesis' field;

(ii) To show that the additional input of human manually made bounding boxes at inference/test time is not necessary for extracting relevant features that characterize actions in still images comparing with approaches with traditional person detection achieving superior mAP performance;

(iii) To tackle the semantic content based image retrieval problem applied to action-based image retrieval;

(iv) To propose a novel action-based image retrieval system without exploiting any additional human-made information in test time but also achieving state-of-the-art performance in mAP when compared to CBIR state-of-the-art works;

(v) To study and introduce an effective (in terms of speed and computational cost) encoding method.

1.3 CONTRIBUTIONS

This thesis contributes to different aspects of still image action recognition and retrieval, which can be used combined or separate in different application scenarios.

The vast scale of scholarly literature, especially in recent years, occasions various problems. One is how to comprehensively record and assess the state of knowledge

on a particular topic. Systematic reviews search, appraise and collate all relevant empirical evidence in order to provide a complete interpretation of research results. Therefore, chapter 3 presents a Systematic Literature Review (SLR) of still image action recognition (SIAR) and of action-based image retrieval mechanisms that considers only still images. The outcome of the SLRs offers insights that could be valuable for computer vision researchers in understanding the various challenges involved in comprehending, identifying, recognizing and retrieving actions in still images.

The semantic gap, especially in action-based image retrieval, has been addressed by only a few works, none of which delve into promising deep learning methods that have greatly improved results in various computer vision tasks. In Chapter 4, the first framework, Act-CBIR, is presented. The main contributions of this chapter are: (i) an efficient action-based image retrieval that relies only on still images for both training and testing. Since the systematic review presented the gap in research in CBIR for actions on still images, this is a relevant research contribution; (ii) it is demonstrated that generic person detectors are suboptimal for the task of characterizing actions in still images, highlighting the efficiency of the newly introduced action detector; (iii) it is shown that the additional input of human manually made bounding boxes at test time is not necessary and can be replaced by detectors capturing human-object interactions. Again, the systematic review presented that most of the efficient still image action recognition systems are dependent on annotations (bounding boxes) in inference/test time, which makes them inefficient for practical applications; (iv) The large amount of data being consumed and generated every day brings the need for agile and efficient search engines. Therefore, the newly introduced encoding method to binarize the feature vector to use the hamming distance fills this gap.

Sequentially, in Chapter 5, the strategy is evolved to obtain features considering multiple inputs to provide context information to the system, and also hint learning approaches, combining higher-level semantic features and local features to obtain a better action representation. The analysis indicates that methods generating multiple regions of interest perform better than those based only on full images. Furthermore, models utilizing multiple information sources outperformed others, emphasizing the significance of incorporating multiple sources for semantic image retrieval. The results were benchmarked against the existing and state-of-the-art works in the image retrieval domain and achieved significantly higher scores. Thus, this work makes a significant contribution to reducing the semantic gap in image retrieval, offering a practical and scalable solution for action-based image retrieval.

1.4 DOCUMENT ORGANISATION

The remainder of this thesis is organized as follows. Chapter 2 provides the necessary theoretical foundation to comprehend the methods developed in this thesis.

It covers computer vision concepts used throughout, including convolutional neural networks and modern convolutional object detectors. Additionally, it explores the fundamentals of still image action recognition and its specifics, along with the theoretical background for content-based image retrieval.

Chapter 3 consists of a Systematic Literature Review (SLR) for still image action recognition and another SLR for action-based CBIR mechanisms.

In Chapter 4, the Act-CBIR method is presented. It is a two-stage image retrieval approach based on action detection and compact codes that eliminate the need for human bounding-box annotations during testing. The method relies solely on still images for both training and testing. The chapter also reports on several experiments using publicly available datasets.

Chapter 5 introduces the Act-Retrieval framework, which is designed for action-based image retrieval using only still images for training and testing. The framework includes modules for action detection, hint-learning, and an attention mechanism to incorporate context information.

Chapter 6 concludes this thesis by summarizing the research's accomplishments and providing perspectives on future work.

2 THEORICAL FOUNDATION

This chapter provides the theoretical foundation necessary to understand what was developed in this thesis. It starts with section 2.1, addressing computer vision concepts that are used throughout this thesis, covering hot topics such as convolutional neural networks and modern convolutional object detectors. In section 2.2, it is presented the fundamentals of still image action recognition along with its specifics, the definition of low-level features and high-level cues, and how authors learn actions in static images. Section 2.3 presents the theoretical background for content-based image retrieval. Finally, this chapter ends with a discussion of the aforementioned topics.

2.1 COMPUTER VISION

This section focuses on the key computer vision concepts, specifically convolutional neural networks and convolutional object detectors, that are crucial for understanding this thesis.

2.1.1 Convolutional Neural Networks

Convolutional Neural Networks are similar to ordinary Neural Networks: they are made up of neurons that have learnable weights and biases (HAYKIN; NETWORK, 2004). Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. The whole network still expresses a single differentiable score function: from the raw image pixels on one end to class scores at the other. And they still have a loss function (e.g. Support-Vector Machines (SVM)/Softmax) on the last (fully-connected) layer.

CNNs take advantage of the fact that the input consists of images and they constrain the architecture in a more sensible way. In particular, unlike a regular Neural Network, the layers of a CNN have neurons arranged in 3 dimensions: width, height, depth (note that the word depth here refers to the third dimension of an activation volume, not to the depth of a full Neural Network, which can refer to the total number of layers in a network). For example, the input images in CIFAR-10 (KRIZHEVSKY; HINTON, G., et al., 2009) are an input volume of activations, and the volume has dimensions $32 \times 32 \times 3$ (width, height, depth respectively). The neurons in a layer will only be connected to a small region of the layer before it, instead of all of the neurons in a fully-connected manner. Moreover, the final output layer would for CIFAR-10 have dimensions $1 \times 1 \times 10$, because by the end of the CNN architecture the full image will be reduced into a single vector of class scores, arranged along the depth dimension (KARPATHY, 2016). Figure 1 summarizes the previous steps and the role of each block will be explained in detail throughout this chapter.

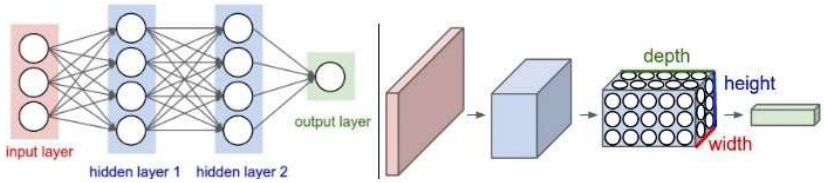


Figure 1 – On the left: a regular 3-layer network. Right: CNN organizes its neurons in three dimensions (width, height, and depth) as viewed in one of the layers. Each layer of a CNN transforms the 3D input volume into a 3D output volume of neuron activations. In this example, the red input layer contains the image, so its width and height would be the dimensions of the image and the depth would be 3 (red, green, and blue channels) (KARPATHY, 2016).

The equation for the neural network (left side in Figure 1) is a linear combination of the independent variables and their respective weights and bias (or the intercept) term for each neuron. In the above neural network, each neuron of the first hidden layer takes as input the three input values and computes its output as follows:

$$y = f\left(\sum_{i=1}^{3} x_i * w_i + b\right) \tag{1}$$

where y is the output vector, x_i are the input values from the ith layer, w_i the weights, b the bias and *f()* an activation function. Then, the neurons of the second hidden layer will take as input the outputs of the neurons of the first hidden layer and so on.

As described above, a simple CNN is a sequence of layers, and every layer of a CNN transforms one volume of activations to another through a differentiable function. Karpathy (2016) stated that three main types of layers are used to build CNN architectures: Convolutional Layer, Pooling Layer, and Fully-Connected Layer (exactly as seen in regular Neural Networks).

An example architecture for a basic CNN used for CIFAR-10 classification consists of the following components: INPUT, Convolutional (CONV), Rectified Linear Unit (ReLU)), Pooling (POOL), and Fully Connected (FC). In more detail:

- INPUT [$32 \times 32 \times 3$] will hold the raw pixel values of the image, in this case, an image of width 32, height 32, and with three color channels R,G,B (Red, Blue and Green).

- CONV layer will compute the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume. This may result in volume such as [$32 \times 32 \times 12$] if we decided to use 12 filters.

- RELU layer will apply an elementwise activation function, such as max(0,x) thresholding at zero. This leaves the size of the volume unchanged ([32 $\times$ 32 $\times$ 12]).
- POOL layer will perform a downsampling operation along the spatial dimensions (width, height), resulting in volume such as [16 $\times$ 16 $\times$ 12].
- FC layer will compute the class scores, resulting in a volume of size [1 $\times$ 1 $\times$ 10], where each of the 10 numbers corresponds to a class score, such as among the 10 categories of CIFAR-10. As with ordinary Neural Networks and as the name implies, each neuron in this layer will be connected to all the numbers in the previous volume.

In this way, classification CNNs transform the original image layer by layer from the original pixel values to the final class scores. Note that some layers contain parameters and others don't. In particular, the CONV/FC layers perform transformations that are a function of not only the activations in the input volume, but also of the parameters (the weights and biases of the neurons). On the other hand, the RELU/POOL layers will implement a fixed function. The parameters in the CONV/FC layers can be trained with gradient descent so that the class scores that the CNN computes are consistent with the labels in the training set for each image (KRIZHEVSKY et al., 2012).

The convolution layer is an initial part of CNN architecture after the input layer consisting of a combination of convolution neurons, also called kernels. Each kernel is associated with a small portion of the image, which is called a receptive field. It operates by dividing the input image into smaller pieces of images and convolving them with a specific set of weights (SALEEM et al., 2022). At ith convolution, we can denote the operations of a convolutional layer as showed in equation 2:

$$
\begin{aligned}
&\forall n \in [1, 2, ..., n_C^l], \\
&Conv(a^{l-1}, K^n)_{x,y} = \varphi^l \left(\sum_{i=l}^{n_H^{l-1}} \sum_{j=l}^{n_W^{l-1}} \sum_{k=l}^{n_C^{l-1}} K_{i,j,k}^n a_{x+i-l,k}^{l-1} + b_n^l \right). \\
&Dim(Conv(a^{l-1}, K^n)) = n_H^l, n_W^l. \\
&\text{Thus,} \\
&a^l = \left[\varphi^l(Conv(a^{l-1}, K^1)), \varphi^l(Conv(a^{l-1}, K^2)), ..., \varphi^l(Conv(a^{l-1}, K^{n_C^{[l]}})) \right]. \\
&Dim.a^l = (n_H^l, n_W^l, n_C^l with n_{H/W}^l = [n_{H/W}^l] + 2p^l - f^l/s^l + 1]; S > 0 \\
&= n_{H/W}^l + 2p^l - f^l; s = 0.
\end{aligned}
\tag{2}
$$

Where the inputs are a^{l-1} with size $(n_H^{l-1}, n_W^{l-1}, n_C^{l-1})$. a^0 is the image input, padding is denoted by p^l, S^l is the stride, the filters are denoted by n_C^l, where each K^n has (f^l, f^l, n_C^{l-1} dimensions, the bias of the nth convolution b_n^l, the activation function is

φ^l and the output a^l with size (n^l_H, n^l_W, n^l_C). The learned parameters at the *lth* layer will be filters with $(f^l * f^l * n^{l-1}_C) * n^l_C$ and bias with $(1 * 1 * 1) * n^l_C$ parameters.

An analogy to explain a convolutional layer is to imagine a flashlight that is shining over the top left of the image. Suppose that the light this flashlight shines covers a 5×5 area. Now, imagine this flashlight sliding across all the areas of the input image. In machine learning terms, this flashlight is called a filter (or sometimes referred to as a neuron or a kernel) and the region that it is shining over is called the receptive field. Now, this filter is also an array of numbers (the numbers are called weights or parameters). A very important note is that the depth of this filter has to be the same as the depth of the input (this makes sure that the math works out), so the dimensions of this filter is $5 \times 5 \times 3$. Now, take the first position the filter is in, for example, it would be the top left corner. As the filter is sliding, or convolving, around the input image, it is multiplying the values in the filter with the original pixel values of the image (computing element-wise multiplications). These multiplications are all summed up (mathematically speaking, this would be 75 multiplications in total). This number is just representative of when the filter is at the top left of the image. This process is repeated for every location on the input volume. The next step would be moving the filter to the right by 1 unit, then right again by 1, and so on. Every unique location on the input volume produces a number. After sliding the filter over all the locations, there will be an array of $28 \times 28 \times 1$ numbers, which is called an activation map or feature map. The reason a 28×28 array is obtained is that there are 784 different locations that a 5×5 filter can fit on a 32×32 input image. These 784 numbers are mapped to a 28×28 array. If two $5 \times 5 \times 3$ filters instead of one are used, then the output volume would be $28 \times 28 \times 2$. By using more filters, we are able to preserve the spatial dimensions better. Mathematically, this is what's going on in a convolutional layer (KARPATHY, 2016).

The convolutional layer's parameters consist of a set of learnable filters. Every filter is small spatially (along width and height), but extends through the full depth of the input volume. For example, a typical filter on the first layer of a CNN might have size $5 \times 5 \times 3$ (i.e., 5 pixels width and height, and 3 because images have depth 3, the color channels). During the forward pass, each filter is slid (more precisely, convolved) across the width and height of the input volume and compute dot products between the entries of the filter and the input at any position. When sliding the filter over the width and height of the input volume, a 2-dimensional activation map that gives the responses of that filter at every spatial position is produced. Intuitively, the network will learn filters that activate when they see some type of visual feature such as an edge of some orientation or a blotch of some color on the first layer, or eventually entire honeycomb or wheel-like patterns on higher layers of the network. Finally, we will have an entire set of filters in each CONV layer (e.g. 12 filters), and each of them will produce a separate two-dimensional activation map, which are stacked along the depth dimension and produce

the output volume (KRIZHEVSKY et al., 2012).

Three hyperparameters control the size of the output volume: the depth, stride, and zero-padding. The first hyperparameter is the depth of the output volume, and corresponds to how many filters are used, each learning to look for something different in the input. For example, if the first convolutional layer receives the raw image as input, then different neurons along the depth dimension may be activated in the presence of several oriented edges or color blobs (KARPATHY, 2016). Second, you must specify the step with which the filter will be slid. When the stride is equal to 1, the filters move one pixel at a time. When it is equal to 2 (or unusually 3 or more, although this is rare in practice), then the filters skip 2 pixels at a time as they are slid. This will produce spatially smaller output volumes (KARPATHY, 2016; KRIZHEVSKY et al., 2012).

According to Karpathy (2016), sometimes it is convenient to add the input volume with zeros around the edge. The size of this zero-padding is a hyperparameter. An interesting feature of zero-padding is that it allows you to control the spatial size of the output volumes.

The spatial size of the output volume (N) can be computed as a function of the input volume size (W), the receptive field size of the Conv Layer neurons (F), the stride with which they are applied (S), and the amount of zero padding used (P) on the border. Equation 3 presents a formula for calculating how many neurons best fit according to Karpathy (2016).

$$N = (W - F + 2P)/S + 1 \tag{3}$$

It is common to periodically insert a pooling layer in-between successive convolutional layers in a CNN architecture. Its function is to progressively reduce the spatial size of the representation to reduce the number of parameters and computation in the network, and hence to also control overfitting. The Pooling Layer operates independently on every depth slice of the input and resizes it spatially, using the MAX operation.

The most common form is a pooling layer with filters of size 2×2 applied with a stride of 2 downsamples every depth slice in the input by 2 along both width and height, discarding 75% of the activations. Every MAX operation would, in this case, be taking a max over 4 numbers (little 2×2 region in some depth slice). That is, the pool layer takes a $k \times k$ region and outputs a single value, which is the maximum in that region, if we consider the maxpooling function. For instance, if their input layer is a $N \times N$ layer, they will then output a $N/k \times N/k$ layer, as each $k \times k$ block is reduced to just a single value via the max function.

Many types of normalization layers have been proposed for use in CNN architectures, sometimes with the intentions of implementing inhibition schemes observed in the biological brain. However, these layers have since fallen out of favor because in practice their contribution has been shown to be minimal if any (KARPATHY, 2016).

For deep CNN models, over-fitting represents the central issue associated with obtaining generalization. Overfitting occurs when the model perform well on training data and does not succeed on test data (unseen data). Various intuitive concepts are used to help the regularization to avoid overfitting, such as dropout, where during each training epoch, neurons are randomly dropped; data augmentation; drop-weights, similar to dropout, in each training epoch, the connections between neurons (weights) are dropped rather than dropping the neurons; and the widely used Batch Normalization, that standardizes the inputs to a layer for each mini-batch.

In the fully connected layer, the neurons have full connections for all the activations in the previous layer. Their activations can hence be computed with a matrix multiplication followed by a bias offset.

Mathematical representations of the fully connected layer are analogous to Equation 1.

Figure 2 presents a generic example of classification through a CNN.

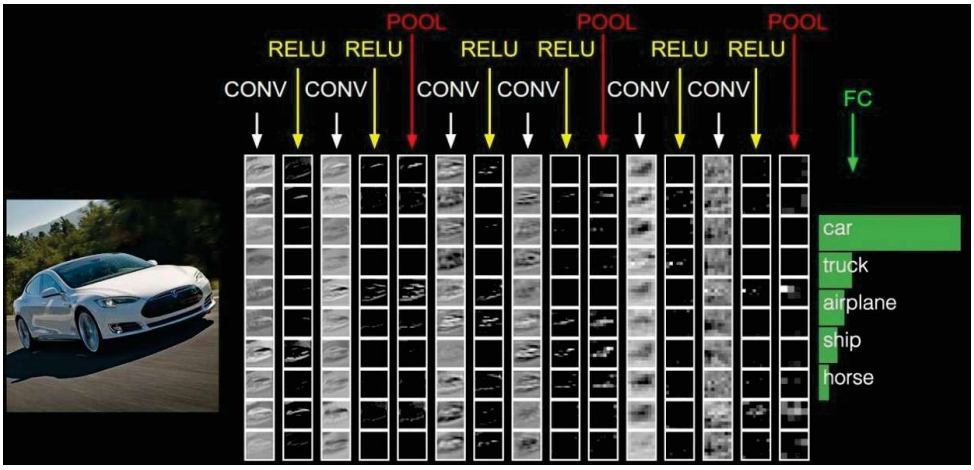


Figure 2 – The initial volume stores the pixels of the raw image (left) and the last volume stores the scores of the class (right). Each volume of activations along the processing path is displayed as a column. Since 3D volumes are difficult to visualize, the slices of each volume are set in rows. The last volume of the layer maintains the scores for each class, but here, we only see the 5 graded scores, and the labels of each of them are displayed (LECUN et al., 2015).

The largest bottleneck to be aware of when constructing CNN architectures is the memory bottleneck (KARPATHY, 2016), because matrix multiplication can grow exponentially.

2.1.1.1 Transfer Learning

Transfer learning is considered as the transfer of knowledge from one learned task to a new task in machine learning (TORREY; SHAVLIK, 2009). In the context of CNN, it is transferring learned features of a pre-trained network to a new problem. Training a convolutional neural network from the beginning in each case usually is not

effective when there is not sufficient amount of training data. The common practice in deep learning for such cases is to use a network that is trained on a large data set for a new problem. While the initial layers of the pre-trained network can be fixed, the last few layers must be fine-tuned to learn the specific features of the new data set. Transfer learning usually results in faster training times than training a new convolutional neural network because you do not need to estimate all the parameters in the new network (TORREY; SHAVLIK, 2009).

In practice, it is unusual to train an entire CNN from scratch (with random initialization), because it is relatively rare to have a dataset of sufficient size as well as computational power. Instead, it is common to pre-train a CNN on a very large dataset (e.g. ImageNet, which contains 1.2 million images with 1000 categories), and then use the CNN either as an initialization or a fixed feature extractor for the task of interest. In such cases, transfer learning can save a significant amount of labeling effort (PAN; YANG, Q., 2009).

The low-level and high-level features learned by a CNN on a source domain can often be transferred to augment learning in a different but related target domain. For target problems with abundant data, we can transfer low-level features, such as edges and corners, and learn new high-level features specific to the target problem (XIE et al., 2015). Deep features extracted from CNNs trained on large annotated datasets of images have been used as generic features very effectively for a wide range of vision tasks (DONAHUE et al., 2014; OQUAB et al., 2014).

2.1.1.2 Modern CNNs

Various modifications have been proposed in the CNN architecture from its earlier days until nowadays. Such modifications include structural reformulation, regularization, parameter optimizations, etc. Conversely, it should be noted that the key upgrade in CNN performance occurred largely due to the processing-unit reorganization, as well as the development of novel blocks. In particular, the most novel developments in CNN architectures were performed on the use of network depth.

The addition of more layers in a CNN makes it capable of learning a large training dataset and efficiently representing more complex mapping functions from inputs to outputs, increasing its capacity. Nevertheless, a problem with training networks with many layers (as deep CNNs) is that the gradient diminishes as it is propagated backward through the network. The error may be so small by the time it reaches layers close to the input of the model that it may have very little effect, making it difficult to know which direction the parameters should move to improve the cost function. This problem is referred to as the “vanishing gradients” problem (GOODFELLOW et al., 2016).

Many fixes and workarounds have been proposed and investigated to deal with this issue, such as alternate weight initialization schemes, unsupervised pre-training,

layer-wise training, and variations on gradient descent, and also archtechtures such as ResNet (HE, K. et al., 2016) which employs skip connections that act as gradient "superhighways", allowing the gradient to flow unhindered.

Another recent trend is the using high-resolution images with CNNs. High-resolution representations are necessary for position-sensitive vision tasks, such as semantic segmentation, object detection, and human pose estimation. In the present up-to-date frameworks, the input image is encoded as a low-resolution representation using a subnetwork that is constructed as a connected series of high-to-low resolution convolutions. The low-resolution representation is then recovered to become a high-resolution one. Alternatively, high-resolution representations are maintained during the entire process using a novel network, referred to as a High-Resolution Network (HRNet) (WANG, J. et al., 2020).

However, even with all these recent advances, building a deep CNN for a specific task highly relies on human expertise, hindering its wide application. Meanwhile, Automated Machine Learning (AutoML) is a promising solution for building a Deep Learning (DL) system without human assistance and is being extensively studied covering data preparation, feature engineering, hyperparameter optimization, and neural architecture search (NAS). NAS aims to search for a robust and well performing neural architecture by selecting and combining different basic operations from a predefined search space (TAN, Mingxing et al., 2019). A review of state-of-the-art AutoML methods is presented by Xin He et al. (2021).

Another recent architecture that is worth mentioning is the EfficientNet family (TAN, Mingxing; LE, 2019). The authors of the EfficientNet family of CNNs (TAN, Mingxing; LE, 2019) developed a new baseline network by performing a neural architecture search using the AutoML MNAS framework, but optimizing FLOPS rather than latency since they state that they were not targeting any specific hardware device. The resulting architecture, EfficientNet-B0, uses Mobile Inverted Bottleneck Convolution, similar to MobileNetV2 and MnasNet, but EfficientNet-B0 is slightly bigger due to the larger FLOPS target. Though EfficientNets perform well on ImageNet, even superpassing state-of-the-art accuracy with up to 10x better efficiency, the authors show that it also transfers well to other datasets, achieving state-of-the-art accuracy on CIFAR-100 and Flowers three other transfer learning datasets, with an order of magnitude fewer parameters.

An strategy that achieves more than 90% top-1 accuracy on ImageNet was recently introduced by Wortsman et al. (2022). Their reasoning is that, unlike like a conventional ensemble strategies, averaging many models without incurring any additional inference or memory costs can be achieved by averaging the weights of multiple models fine-tuned with different hyperparameter configurations, improving accuracy and robustness. Petersen et al. (2022) proposes a loss function for classification aiming

to train the network while not only considering the only the top-1 prediction, but also, e.g., the top-2 and top-5 predictions. This introduced function is called differentiable top-k cross-entropy, where k is a positive integer, such as 1 or 5, leading to top-1 or top-5 training objectives. Their evaluation concludes that that relaxing k does not only produce better top-5 accuracies, but also leads to top-1 accuracy improvements.

In this subsection, it was presented several recent advances in CNNs. A deeper review on modern CNN architectures and advances can be found at Asifullah Khan et al. (2020), Zewen Li et al. (2021), Lindsay (2021) and Ghimire et al. (2022), although new and promising technologies on this topic are presented almost daily.

2.1.2 Object Detectors

Object detection deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of object detection include face detection and pedestrian detection. Object detection has applications in many areas of computer vision, including image retrieval and video surveillance.

One brute force approach for object detection is to slide windows from left and right, and from up to down to identify objects using classification. To detect different object types at various viewing distances, windows of varied sizes and aspect ratios are used. Figure 3 presents the slide window mechanism for object detection.

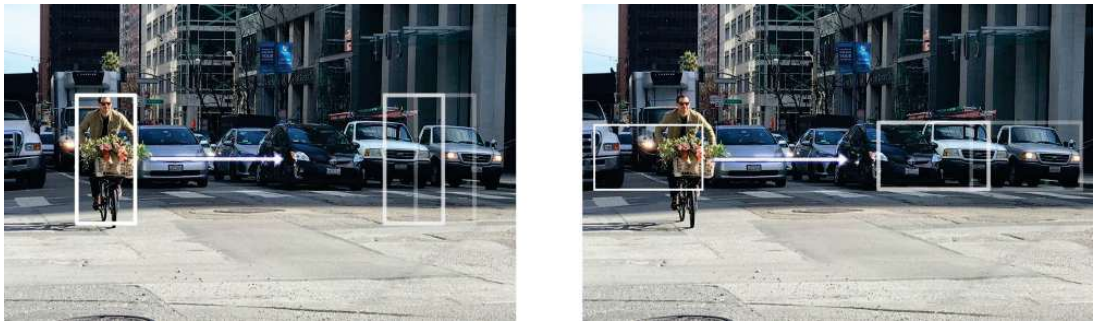


Figure 3 – Slide window mechanism for object detection.

Patches are cut from the picture according to the sliding windows. These patches are warped since many classifiers take fixed size images only. The warped image patch is fed into a CNN classifier to extract features. A SVM classifier is applied to identify the class and another linear regressor for the boundary box.

Since AlexNet won the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) challenge, the use of the CNN has dominated the field. In the next subsection some of the highlights of this literature are survey.

2.1.2.1 R-CNN

The R-CNN paper by Girshick et al. (2014) was among the first modern convolutional neural network based detection approach.

R-CNN is a state-of-the-art region-based visual object detection system that combines bottom-up region proposals with rich features computed by a convolutional neural network. At the time of its release, R-CNN improved the previous best detection performance on PASCAL VOC 2012 by 30% relative, going from 40.9% to 53.3% mean average precision. Unlike the previous best results, R-CNN achieves this performance without using contextual rescoring or an ensemble of feature types (GIRSHICK et al., 2014).

Instead of a brute force approach, R-CNN uses a region proposal method to create ROIs for object detection. The regions are warped into fixed size images and fed into a CNN network individually. It is then followed by fully connected layers to classify the object and to refine the boundary box.

R-CNN method took the straightforward approach of cropping externally computed box proposals out of an input image and running a neural net classifier on these crops. However, this approach can be expensive because many crops are necessary, leading to significant duplicated computation from overlapping crops.

Figure 4 presents a schematic representation of this network.

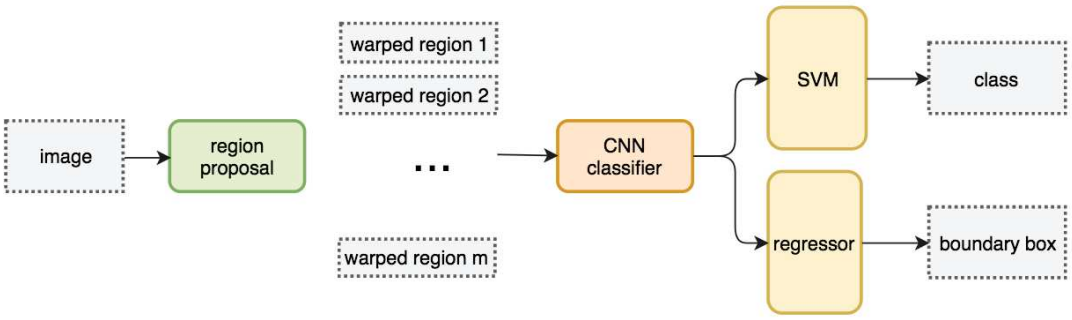


Figure 4 – R-CNN network flow. It applies region proposal on feature maps and form fixed size patches using ROI pooling.

2.1.2.2 Fast R-CNN

Fast R-CNN (GIRSHICK, 2015) alleviated this problem by pushing the entire image once through a feature extractor then cropping from an intermediate layer so that crops share the computation load of feature extraction.

Fast R-CNN also uses an external region proposal method, like selective search, to create ROIs which later combine with the corresponding feature maps to form patches for object detection. The patches are warped to a fixed size using ROI pooling and fed to fully connected layers for classification and localization (detecting the location of the

object). By not repeating the feature extractions, Fast R-CNN cuts down the process time significantly. Figure 5 summarizes the Fast R-CNN Network flow.

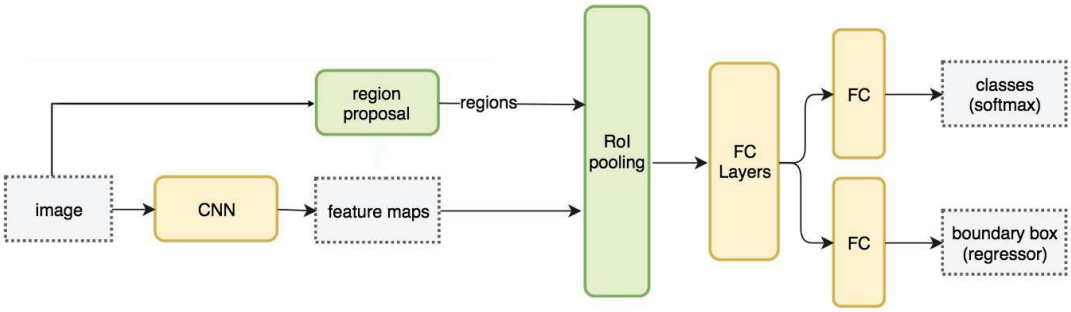


Figure 5 – Fast R-CNN Network flow.

Fast R-CNN depends on an external region proposal method like selective search. However, those algorithms run on CPU and they are slow. In testing, Fast R-CNN takes 2.3 seconds to make a prediction in which 2 seconds are for generating 2000 ROIs.

2.1.2.3 Faster R-CNN

Faster R-CNN (REN et al., 2015) adopts similar design as the Fast R-CNN, except it replaces the region proposal method by an internal deep network and the ROIs are derived from the feature maps instead. The new Region Proposal Network (RPN) is more efficient and runs at 10ms per image in generating ROIs.

The network flow is similar but the region proposal is now replaced by a convolutional network (RPN).

In the Faster R-CNN setting, detection happens in two stages. In the first stage, called the region proposal network, images are processed by a feature extractor (e.g., VGG-16), and features at some selected intermediate level (e.g., “conv5”) are used to predict class-agnostic box proposals. The loss function for this first stage takes form using a grid of anchors tiled in space, scale and aspect ratio. In the second stage, these (typically 300) box proposals are used to crop features from the same intermediate feature map which are subsequently fed to the remainder of the feature extractor in order to predict a class and class-specific box refinement for each proposal. The loss function for this second stage box classifier also takes form using the proposals generated from the RPN as anchors. Notably, one does not crop proposals directly from the image and re-run crops through the feature extractor, which would be duplicated computation. However there is part of the computation that must be run once per region, and thus the running time depends on the number of regions proposed by the RPN (REN et al., 2015).

2.1.2.4 R-FCN

Fast and Faster R-CNN methodologies consist in detecting region proposals and recognizing an object in each region. The Region-based Fully Convolutional Network (R-FCN) released by (DAI et al., 2016) is a model with only convolutional layers allowing complete backpropagation for training and inference. The authors have merged the two basic steps in a single model to take into account simultaneously the object detection (location invariant) and its position (location variant).

A ResNet-101 CNN (SZEGEDY et al., 2017) model takes the initial image as input. The last layer outputs feature maps, each one is specialized in the detection of a category at some location. For example, one feature map is specialized in the detection of a cat, another one in a banana and so on. Such feature maps are called position-sensitive score maps because they take into account the spatial localization of a particular object. It consists of k×k×(C+1) score maps where k is the size of the score map, and C the number of classes. All these maps form the score bank. Basically, R-FCN creates patches that can recognize part of an object. Figure 6 summarizes the R-FCN pipeline considering a ReNet as backbone.

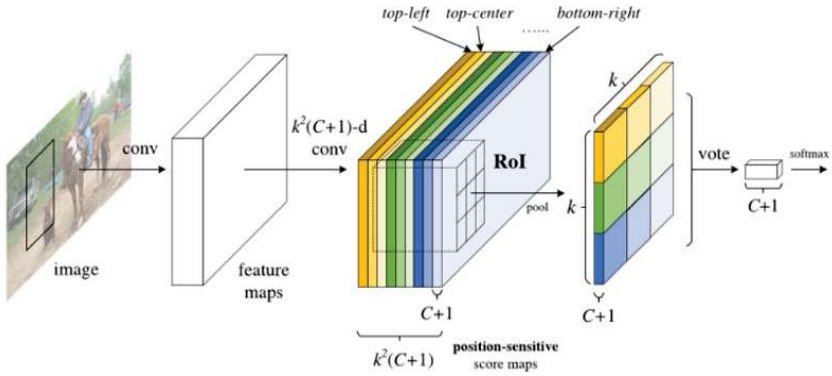


Figure 6 – The input image feeds a ResNet model to produce feature maps. A RPN model detects the Region of Interests and a score is computed for each region to determine the most likely object if there is one (DAI et al., 2016).

In parallel, R-FCN runs a RPN to generate regions of interest. Finally, each ROI is cut in bins and checked against the score bank. If enough of these parts are activated, then the patch vote 'yes', and the object is recognized.

2.1.2.5 YOLO

Faster R-CNN has a dedicated region proposal network followed by a classifier. Region-based detectors are accurate but not without a cost. Faster R-CNN processes about 7 Frames per second (FPS) for PASCAL VOC 2007 testing set. Like R-FCN, researchers are streamlining the process by reducing the amount of work for each ROI.

You-only-Look-Once (YOLO) (REDMON et al., 2016) uses a single convolutional network that simultaneously predicts multiple bounding boxes and class probabilities for

those boxes. YOLO trains on full images and directly optimizes detection performance. This unified model has several benefits over traditional methods of object detection, being extremely fast, accurate and capable of learning generalizable representations of objects.

Figure 7 shows a schematic representation of this network.

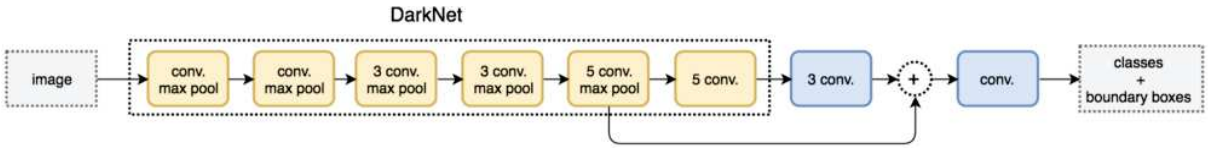


Figure 7 – YOLO Network flow.

However, it does not make independent detections using multi-scale feature maps. Instead, it partially flattens features maps and concatenates it with another lower resolution maps. For example, YOLO reshapes a $28 \times 28 \times 512$ layer to $14 \times 14 \times 2048$. Then it concatenates with the $14 \times 14 \times 1024$ feature maps. Afterward, YOLO applies convolution filters on the new $14 \times 14 \times 3072$ layer to make predictions.

In its recent fourth version, YOLOv4 (BOCHKOVSKIY et al., 2020) achieves state-of-the-art results at a real-time speed on the Microsoft COCO dataset with 43.5% AP running at 65 FPS on a Tesla V100. Its architecture is composed of CSPDarknet53 as a backbone, spatial pyramid pooling additional module, PANet path-aggregation neck, and YOLOv3 head (REDMON; FARHADI, 2018). The authors present a series of methods to improve the object detector's accuracy divided between two categories: Bag of freebies (BoF), methods to improve accuracy without increasing the inference cost, i.e., they act on the training strategies, such as data augmentation strategies, Complete Intersection over Union loss and DropBlock regularization among others; and Bag of specials (BoS), methods that increase the inference cost by a small amount but can significantly improve the accuracy, such as Spatial pyramid pooling, Mish activation, and a modified Spatial Attention Module, among others.

2.1.2.6 SSD

Single Shot MultiBox Detector (SSD) (LIU, W. et al., 2016) discretizes the output space of bounding boxes into a set of bounding box priors over different aspect ratios and scales per feature map location. At prediction time, the network generates confidences that each prior corresponds to objects of interest and produces adjustments to the prior to better match the object shape. Additionally, the network combines predictions from multiple feature maps with different resolutions to naturally handle objects of various sizes.

Though the SSD paper was published only recently (LIU, W. et al., 2016), the term SSD is used to refer broadly to architectures that use a single feedforward convolutional network to directly predict classes and anchor offsets without requiring a second

stage perproposal classification operation. Figure 8 presents the general structure of the SSD detector using a VGG-16 as backbone.

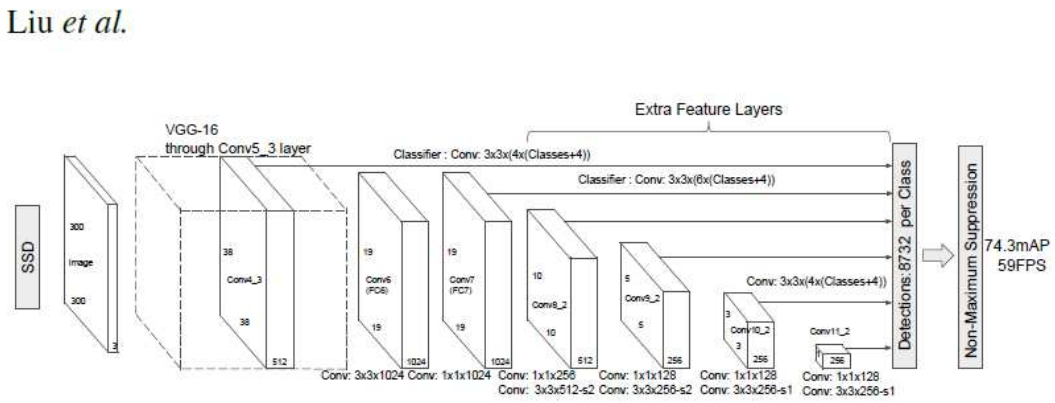


Figure 8 – The SSD model adds several feature layers to the end of a base network (backbone), which predict the offsets to default boxes of different scales and aspect ratios and their associated confidences (LIU, W. et al., 2016).

2.2 STILL IMAGE ACTION RECOGNITION

A still image action recognition (SIAR) system usually follows a certain set of steps, presented in Figure 9

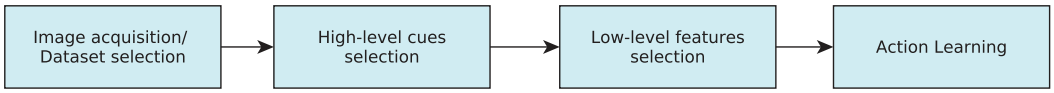


Figure 9 – Still image action recognition flowchart.

In still image-based action recognition, usually, the low-level features extracted directly from the whole image may not work well. Thus, previous works seldom rely only in the whole image or scene for low-level feature extraction and action recognition.

Since only the spatial information is available in single images with a cluttered background, researchers have pursued different high-level cues in still images in order to characterize actions instead of using low-level features in the whole image.

The high-level cues can be characterized through several low-level features. Then, different high-level cues can be combined to recognize actions in still images. This section presents the high-level cues and low-level features that have been used for still-image-based action recognition.

2.2.1 High-level cues

The most popular high-level cues for still-image-based action recognition include parts of the image that are thought to be most important for recognition, such as human

body, body parts, action-related objects, human-object interactions, and the whole scene or context. These cues can characterize human actions from different aspects.

The human body is an important cue for still-image-based action recognition. Most of the existing approaches use the human body cue for action representation (GUO, G.; LAI, 2014; ZHU, F. et al., 2016). It can be detected automatically in images or manually labeled. Usually, the bounding box of the human is used to indicate the location of the person and determine the image region for human body information extraction.

Rather than the whole human body, body parts can be more related to action execution. When performing different actions, e.g., throwing a ball or using a computer, the body parts (in this case, the arms), can be in different locations or with different poses. Based on this, the cue of body parts may be used for action characterization.

When humans perform actions, there are objects related to these actions. Thus it is natural to consider the related objects for human action characterization. Different actions might be related to different objects. By knowing the related objects, it can help to recognize the corresponding actions. For example, a baseball bat is highly possibly related to the action of playing baseball (Figure 10). This way, researchers have realized the importance of using object information to help the action recognition in still images (GUO, G.; LAI, 2014).

Figure 10 – The presence of the baseball bat helps to recognize the action (BARBEE, 2018).

Another important cue is the interaction between humans and objects. For instance, the relative position between a person and the action related object (e.g., a book for reading), and the relative angle between the person and the object (e.g., the person is above the bike when he/she is riding a bike), the relative size of the person and the object (e.g., a phone in calling is much smaller than a horse in riding in the two

different actions), etc.

2.2.2 Low-level features

Above, it was introduced various high-level cues for action analysis in still images. After selecting these high-level cues, they are usually characterized by using low-level features. Different low-level features have been attempted in previous approaches. The most common low-level features for still image-based action recognition will be presented in this section.

One approach to working with image matching is to use local descriptors to represent an image or an image region. Descriptors are feature vectors for an image or certain regions of an image and can be used to compare regions in different images. This low-level feature vectors are usually formed by local or global descriptors. Local descriptors computed at points of interest have proven to be successful in applications such as matching and image recognition. Descriptors are distinct, robust to occlusion and do not require segmentation. The simplest descriptor is a vector with the intensities of the pixels in the image. The cross-correlation measure can then be used to compute the similarity between two regions. However, the high dimensionality of such a descriptor increases the computational complexity of the comparison (MIKOLAJCZYK; SCHMID, 2005).

Typical low-level features include the Scale-invariant Feature Transform (SIFT) and its variant Dense SIFT (DSIFT), Histogram of Oriented Gradients (HOG), Shape Context (SC), gist descriptor, or some other features. It is important to highlight that Convolutional Neural Networks are often used as feature extractors.

An often-used low-level feature descriptor is the Histogram of Oriented Gradients (HOG) descriptor, proposed by Dalal and Triggs (2005). HOG represents an image as a feature vector. For this, the image is divided into small cells with a size of 8 $\times$8 pixels, in each of which a gradient histogram is calculated. For the descriptor to be independent of light variations, the authors propose to normalize the contrast in larger areas (called blocks), which in this case are 16$\times$16 pixels, this result helps to normalize the block cells (each block will contain 4 cells). Each of these blocks is called HOG descriptor. With this, each input image with a size of 144 $\times$ 48 pixels will have a HOG vector of 3060 features.

SC was proposed by (BELONGIE et al., 2001) for shape feature extraction for object matching. The SC can also be used for action recognition in still images. It can help to detect and segment the human contour. The usage of the SC feature is crucial for high-level cue representation of human body silhouettes for action recognition.

Spatial envelop, also called GIST, was proposed by (OLIVA; TORRALBA, 2006). A set of holistic, spatial properties of the scene can be computed by the GIST method. The GIST is an abstract representation of the scene that spontaneously activates

memory representations of scene categories (a city, a mountain, etc.). The GIST feature is mainly used to integrate scene or background information.

Another popular descriptor is the Scale Invariant Feature Transform (SIFT) (LOWE, 1999), where key locations are selected at maxima and minima of a difference of Gaussian function applied in scale space. They can be computed by building an image pyramid with resampling between each level. Furthermore, SIFT locates key points at regions and scales of high variation, making these locations particularly stable for characterizing the image. The authors have demonstrated the stability of SIFT keys to image transformations. For keypoint descriptor creation, a 16×16 neighbourhood around the keypoint is taken. It is divided into 16 sub-blocks of 4×4 size. For each sub-block, 8 bin orientation histogram is created. So a total of 128 bin values are available. It is represented as a vector to form keypoint descriptor.

Due to its superior performance demonstrated in broad visual recognition tasks, most current works employ features learned by convolutional neural networks, explained in more detail in section 2.1.1.

In addition to the frequently used descriptors mentioned above, there are also other low-level features for action recognition, such as Deep Belief Network (DBN) (HINTON, G. E., 2009), Massachusetts Institute of Technology (SURF) (BAY et al., 2006), Circular Histogram of Oriented Rectangles (CHORs) (IKIZLER; DUYGULU, 2009) and several pure color descriptors (VAN DE WEIJER; SCHMID, 2006) (KHAN, F. S. et al., 2012): RGB descriptor, C descriptor, Hue Saturation descriptor, Robust Hue descriptor (HUE), Opponent Derivative descriptor (OPP), Color Name, among others.

2.2.3 Action Learning

Given various image representations, the selected high-level cues represented by low-level features, the next step is to learn the actions from training examples. The learned models or classifiers can then be used to recognize actions from the unseen, test images. Different learning methods have been proposed by researchers.

One of the crucial factors for a successful image classification system is the classifier. A well-designed classifier would not be sensitive to some of the other factors, such as feature extraction. In the past several decades, artificial neural networks benefit a lot from randomly generated parameters, not only in the learning speed but also in the generalization performance (CAO et al., 2016).

For image classification, another important problem is to find a classifier with good generalization. There have been various classifiers, both linear and nonlinear, such as Artificial Neural Network (ANN), support vector machines (SVM), polynomial classifier, and fuzzy rule-based systems.

The action learning methods can be categorized into different categories, such as general models, discriminative learning, learning mid-level features, fusing multiple

features, extracting spatial saliency, conditional random field, and pose matching (GUO, G.; LAI, 2014). They also can be combined to improve the recognition accuracy.

- **Generative models**: Generative models usually learn the statistical distributions for action classes, which can randomly generate the observable data.
- **Discriminative learning**: Discriminative learning is appropriate for distinguishing different action classes, without turning to learning the complex generative models. Discriminative models are a class of models used for modeling the dependence of unobserved (target) variables y on observed variables x. Within a probabilistic framework, this is done by modeling the conditional probability distribution P(y|x), which can be used for predicting y from x.
- **Learning mid-level features**: Different from low-level features such as SIFT and HOG, some middle-level features can be learned from the action images. Most of them are based on the extracted low-level features.
- **Multiple features fusion**: Multiple features can be extracted to help improve the action recognition accuracy, in feature level (e.g., histogram concatenation) or score level. The assumption in fusion-based approaches is that multiple features may complement each other and a combination of them may characterize the actions better than each single feature. Thus the fusions of multiple features are expected to improve the action recognition accuracies.
- **Pose matching**: Some approaches to action recognition are mainly based on matching human body poses. The matching scheme is especially to exploit body shape and pose information. From a sketch of human body poses, it was assumed that there is a great similarity between intra-class poses and the matching of poses can recognize actions. To deal with variations of intra-class poses, multiple poses can be used to represent a certain class.

Convolutional Neural Networks are a special topic in still image action recognition because they can be used for high-level cues selection, as feature extractor, and also as a mechanism for action learning.

2.3 CONTENT-BASED IMAGE RETRIEVAL

The rapid growth of the mobile device market contributed significantly to the increase in the volume of images being generated and consumed each day, and consequently, searching for images in a large database that match a query becomes an increasing necessity. In this context, there are three main types of image search engines: meta-data search, content-based search, and a hybrid approach of the previous

two. Meta-data search is similar to keyword-based search engines and usually does not examines the content of the image, relying on textual cues, such as tags and contextual cues that appear near the image. This additional information may not always be available, which weakens this approach. Most pictures do not have meta-data (as tags or contextual cues), so image search engines that quantity the content of an image, as the Content-Based Image Retrieval (CBIR), are ideal for many applications. Considering the recent success in applying convolutional neural networks (CNNs) based methods in the image domain, many approaches have been proposed in the CBIR context. Even though it is in a relatively young age, CBIR has been an active field of study in recent years due to its nearly endless potential for applications, such as person re-identification, remote sensing, medical image search, shopping recommendation in online markets, and many others (ZHOU, W. et al., 2017; CHEN et al., 2021).

Intuitively, one could think that a image representation based on the L2 normalization of the concatenation of the vector corresponding of the three color channels could provide a good distance metric between a two images. But as we can see in Figure 11, that is not always the case.

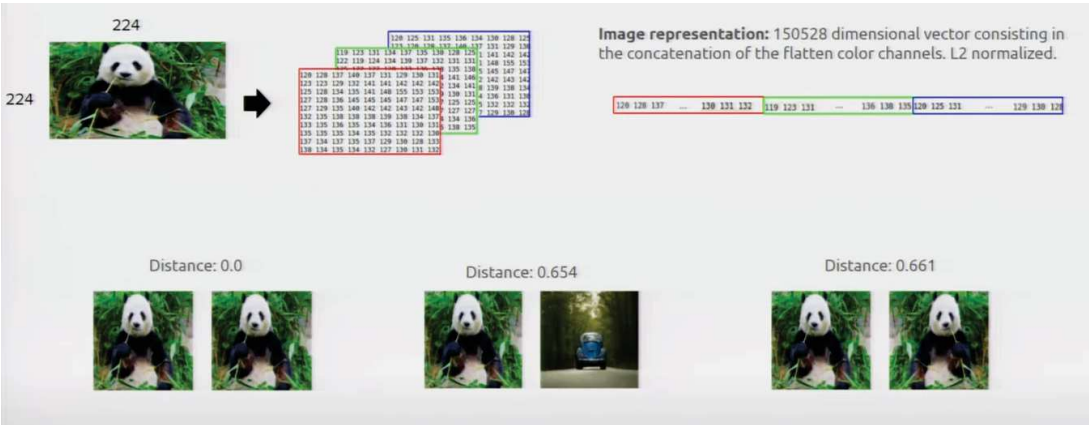


Figure 11 – The challenge of measuring similarity between two images: a simple concatenation of the three image vector of the color channels is not a good similarity metric (WANG, L., 2016).

To further address this issue, researches proposed to compute the similarity between images by the similarity between feature vectors that should comprehend information in terms of the color, texture, shape, gradient, etc. are represented in the form of a feature descriptor. Thus, the performance of any CBIR method heavily depends upon the feature descriptor that should represent the image. In the earlier days in the CBIR research, the feature descriptor representation would utilize the visual cues of the images selected manually based on the need. These approaches, such as SIFT descriptor, are also termed as the hand-designed or hand-engineered feature description (CHEN et al., 2021). Figure 12 presents the general pipeline of a classical CBIR system.

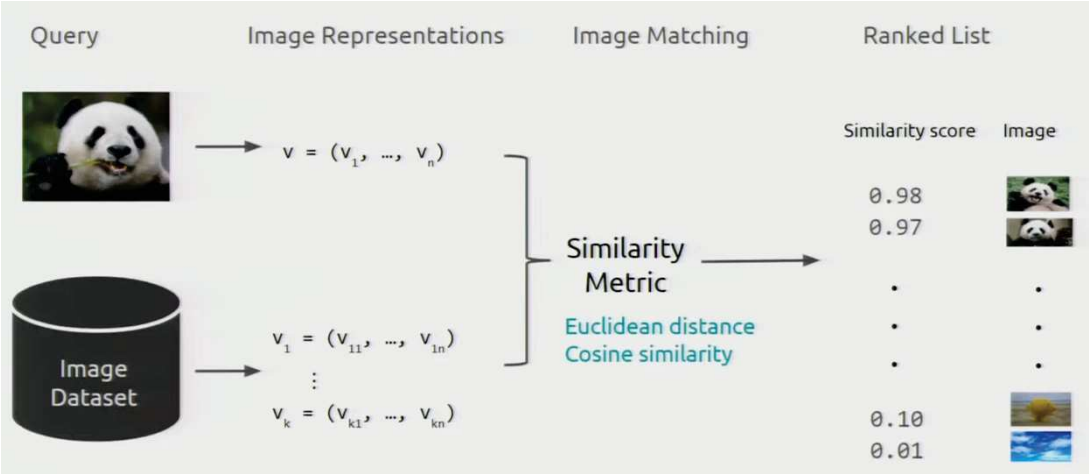


Figure 12 – Classical pipeline for a CBIR system (WANG, L., 2016).

In the past decade, remarkable progress has been made in image feature representations. Due to the success of deep learning for image representation in several tasks, researchers have been trying to apply this hierarchical feature representation technique to learn abstract features from data. The works of Babenko et al. (2014) and Wu and Xiao (2018) show that multilayer representations generated by CNNs can be used to represent images in CBIR for fast and efficient image retrieval solutions.

To form more discriminating image descriptors, further research efforts on indexing algorithms and structures are expected in order to effectively associate low-level features with further semantic information. However, describing images in higher semantics usually leads to generate high-dimensional image features which predominantly have a sparse distribution of data. This obviously will degrade the retrieval performance of CBIR systems, but despite the considerable volume of algorithms conducted for dimensionality reduction, this problem still challenges the CBIR field. The goal here is to project data into another space to highlight certain structures by identifying only the interested projections. Many methods have been proposed in the literature, such as Principal Component Analysis (PCA), Bag of Words (BOW), Fisher Vectors (FV), and Vector of Linearly Aggregated Descriptors (VLAD), and even discussed by some comprehensive studies (ZHUO et al., 2014; ALZU'BI et al., 2015).

To overcome this issue, there is a growing trend in studying deep hashing methods for CBIR, where hash functions and binary codes are learnt using deep convolutional neural networks and then the binary codes can be used to do approximate nearest neighbor search. The main idea of hashing based ANN methods is to map images into a similarity preserved hamming space where the search space can be efficiently pruned (CAI et al., 2017). The hash-based scheme maps the search-key values with a collection of buckets so that their mapped values are determined by a function called the hashing function. Several approaches also have been inserted in literature, such as LSH, Deep pairwise-supervised hashing (DPSH), Semi-supervised

deep hashing (SSDH), and many more (ALZU'BI et al., 2015; DUBEY, 2021). A survey on the theme is presented by Cai et al. (2017).

In general, the structure of feature vectors selected determines the type of distance measure that will be used to compare their similarity. The distance measure mathematically indicates the similarity between the query and each image in the database according to necessity. To achieve more accurate retrieval and better performance, CBIR systems try to employ effective similarity matching measure which accurately characterizes and quantifies the perceptual similarities. Despite the success of utilizing the common distance measures in the literature, such as Minkowski distance, Mahalanobis distance, Cosine distance, Earth Mover's distance, Hamming distance, and Kullback-Leibler and Jeffrey divergence distance, finding an adequate and robust distance measure is still one of the challenging issues in the field of CBIR (DUBEY, 2021).

Moreover, to enhance the performance of CBIR systems, several authors also have proposed pre-processing and post-processing techniques. Pre-processing techniques, such as Direct Resize, TenCrop, PadResize, other data augmentation strategies and semantic distribution bias in datasets, and post-processing methods, such as k-reciprocal, and query expansion have shown promising results(HU et al., 2020).

Despite all these advances, semantic content-based image retrieval is still a challenge. Humans can describe and interpret image easily, including its overall topology and objects using high-level semantic concepts. Unlike humans, digital machines can provide fewer semantic words for the same image. Machines deal with low-level features extracted from image pixels, which provides a numerical description of images, but with a wide gap compared to the human interpretation of the same image. This gap between the richness of high-level human's perception and low-level machine's descriptions is known as the "semantic gap". The user looks up for semantic similarity, but the database can only provide similar images by a digital processing (ALZU'BI et al., 2015). In addition, the semantic gap between image properties and object properties broadly limits the retrieval efficiency as there is inconsistency in understanding visual data for different users (YUE et al., 2011).

The research trend in image retrieval suggests that the deep learning based models are driving the progress. The future work appears to be the exploration of improved deep learning models, more relevant objective functions, minimum loss based quantization techniques, semantic preserving feature learning, and attention focused feature learning. To deal with the semantic gap, attention mechanisms have been observed as a very effective way of modelling the saliency information into the feature space to avoid the effect of background and can boost the performance of image retrieval models significantly. In order o perform the faster image search, the learnt feature or hash code should be as low dimensional and compact as possible. Thus,

better strategy for feature quantization and maximizing the relevant information into feature space in a compact way can be seen as one of the future directions (DUBEY, 2021).

2.4 DISCUSSION

This chapter has presented the main basis of Still Image Action Recognition and Content-based Image Retrieval.

Still Image Action Recognition is relatively less studied and complex problem when compared to action recognition with the aid of video information. This chapter have introduced different approaches based on categorization of high-level cues and low-level features, and also action learning methods have been discussed too. The main issues include occlusions and a lack of movement cues, which may make it difficult to extract the related high-level cues and low-level features.

The image retrieval field has experienced an astonish evolution in recent year using promising deep learning approaches together with the exploration of more relevant objective functions, quantization techniques, semantic preserving feature learning, and attention focused feature learning can boost results especially for fine-grained image retrieval, such as the still image action-based retrieval. Nevertheless, semantic content-based image retrieval is still a challenge. It is difficult to translate low-level features extracted from the image's pixels, which provide a numerical description of the images, to human interpretation of the same image.

As a relatively new area, the researches on still image-based action recognition and semantic CBIR are in their early stage, with researchers exploring promising trends.

3 SYSTEMATIC LITERATURE REVIEW

This chapter presents Systematic Literature Reviews for still image action recognition and for action-based CBIR mechanisms.

3.1 SLR ON SIAR

This research has been carried out using the guidelines proposed by Barbara Kitchenham (2004) to perform Systematic Literature Review (SLR) or Systematic Review (SR), which involves several activities, such as the development of the review protocol, the identification and selection of primary studies, data extraction and synthesis, and reporting the results. These steps were followed for the reported study, as described in the following sections of this document.

The general objective of this first review is to answer the following Research question (RQ):

> RQ. What is the research status on recognizing actions using exclusively static images?

More specifically, this study addresses the following three issues:

- RQ1. What are the main techniques used for recognizing the actions?
- RQ2. What is the degree of accuracy compared to other visual recognition tasks?
- RQ3. Which high-level cues have the greatest impact on action recognition?

3.1.1 Data sources and search strategies

Only documents written in english and available online were searched. The search strategy included the following electronic databases:

- IEEEXplore (www.ieeexplore.ieee.org/Xplore/).
- Scopus (https://www.scopus.com).
- ACM Digital library (www.portal.acm.org/dl.cfm)
- Springer (https://link.springer.com/)
- Biblioteca Nacional de Teses e Dissertações (https://bdtd.ibict.br/vufind/)

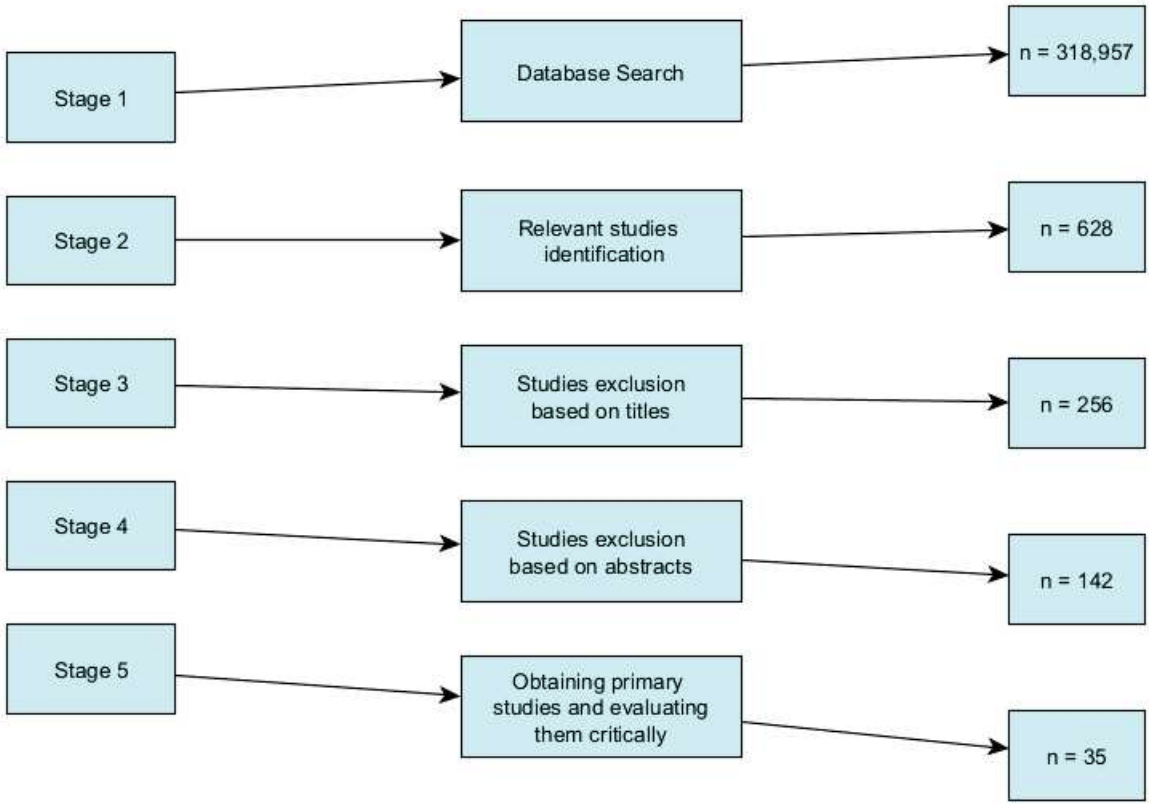


Figure 13 – Stages of the study selection process.

Figure 13 presents the systematic review process and the number of documents identified in each stage.

In phase 1, the terms listed in Table 1 were searched.

Table 1 – Search terms used in this review

Term 1	Term 2	Term 3	Term 4
Still	image	action	recognition
"still image"	action	recognition	-
"static image"	action	identification	-
static	image	action	identification

Editorials, prefaces, article summaries, interviews, news, reviews, correspondences, discussions, comments, readers' letters and tutorial summaries, workshops, panels, and poster sessions were excluded from the search. This research strategy generated a total of 318,957 results that included 628 non-duplicate citations. Figure 14 presents a visualization of the bibliometric network based on terms relations (keywords) considering at least 5 occurrences obtained through the VOSviewer [1] software combining each database.

[1] https://www.vosviewer.com/

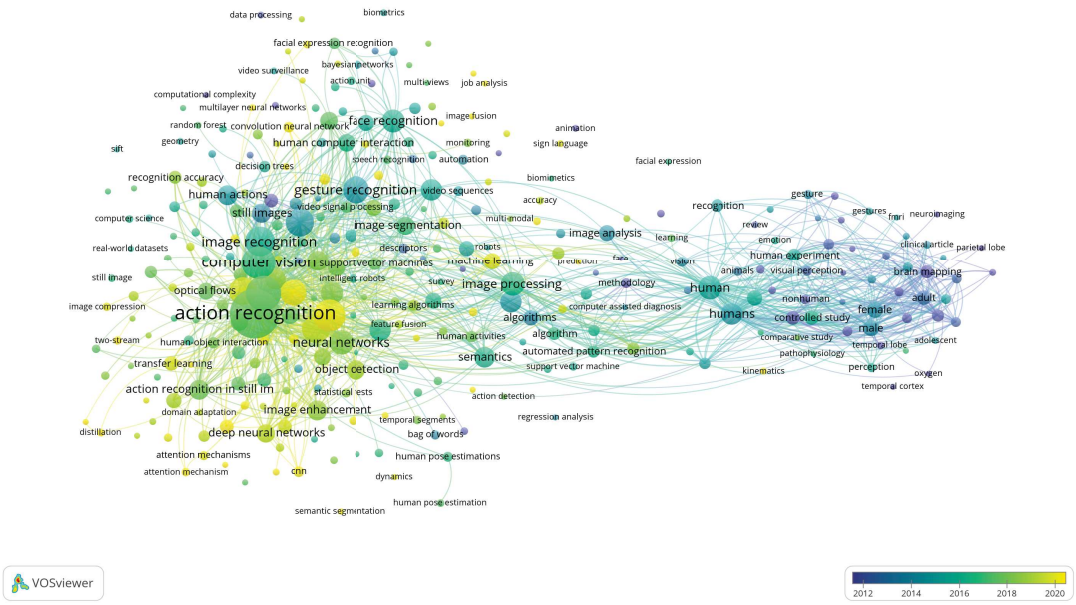


Figure 14 – Image representing the bibliometric network based on terms relations.

It is possible to conclude that research in SIAR is closely linked with scene understanding, which is one of the main objectives of computer vision. We can also note an increase in techniques based on deep learning in recent years.

3.1.2 Managing studies and inclusion decisions

The study followed the citation management procedure reported by Dybå and Dingsøyr (2008). Mendeley was used to store relevant stage 2 citations (n = 628). The citations were then imported into a Microsoft Excel spreadsheet where the sources of each citation and the subsequent inclusion/exclusion decision were recorded.

In the third stage, the titles of all studies resulting from the previous stage were analyzed to determine their relevance to the systematic review. At this stage, articles with titles that clearly indicated that they were outside the scope of the SLR limit were excluded, leaving 256 studies identified as relevant. However, an article's title does not always represent the document content. During the next stage, each abstract was revised to improve the definition of relevance. At the end of stage 4, 142 documents remained for stage 5 of the selection process.

It is worth mentioning that a total of 22 brazilian studies were found in the National Library of Theses and Dissertations. Six of these studies were classified as relevant. However, all these studies dealt with videos or sequences of images for the recognition of actions, and none of them used only still images for action recognition, evidencing the novelty of the topic, especially at the national level.

3.1.3 Final selection

Each of these 142 studies that remained after stage 4 was assessed by the author according to five criteria. These criteria were proposed on the basis of the Critical Appraisal Skills Program (CASP)[2] and the principles of good practices for conducting empirical research in software engineering (KITCHENHAM, B. A. et al., 2002). The five criteria addressed three key quality issues that should be considered in the evaluation of the studies identified in the review:

- **Rigor**: Has a complete and appropriate approach been applied to the main research methods in the study?
- **Credibility:** Are the findings well presented and meaningful?
- **Relevance:** How useful are the findings for the software industry and the research community?

Thus, the following selection criteria were used to ensure that the documents address the research topic.

1. Does the system only relies on still images?
2. Is the purpose of the article clearly stated?
3. Does the document adequately discuss the contextual and technical factors used?
4. Is the proposal tested on public domain datasets?
5. Is the performance obtained close to or better than the state-of-the-art at the date of publication?

These five points provided a measure in which the author is confident that a selected paper could be a valuable contribution to understand and model the problem. Each of the five criteria was classified on a dichotomous scale (“yes” or “no”).

A total of 35 documents were selected from the 142 articles, 35 of which were primary studies and 3 secondary studies. The quality evaluation was performed based on the selection criteria presented. Documents that met the five criteria proposed above were accepted.

The primary and secondary studies selected are enlisted in Appendix A.

[2] http://www.casp-uk.net/

3.1.4 Data extraction and synthesis

From the selected studies after the previous stages, the data were extracted using a predefined data extraction form, as presented in Appendix C. This form allowed to record complete details of the papers being analyzed and specify how each one approached the research questions.

The objectives, settings, descriptions of research methods, findings and conclusions, as reported by the primary studies authors were copied verbatim in Microsoft Excel.

The data was synthesized by identifying themes emanating from the findings reported in each of the paper reviewed in this study. The following section presents frequencies of the number of times each theme is identified in different studies. The respective frequencies reflect the number of times a particular challenge has been mentioned in different papers.

3.1.5 Results

Thirty-five studies were selected for this SLR, 35 primary and 6 secondary (according to Appendix A). These studies covered a number of techniques, were done with a multiplicity of search methods and were performed on different datasets.

The next subsections describe the characteristics of the studies, the applied research methods, and evaluate the quality of the studies according to the proposals of this work.

3.1.5.1 Study overview

Table 2 shows that the number of works on recognition in still images are increasing in recent years.

Table 2 – Selected papers by date of publication

2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
2	1	2	2	1	1	5	3	3	3	6	1	4

It can be argued that the publication trend may be an indicator of professionals and researchers growing interest on the subject, in addition to its obvious complexity. It is verified that in last year there is a greater number of published works.

Compared to video-based action recognition, still image action recognition is less studied, which reveals the difficulty of the problem.

3.1.5.2 High-level Cues

In action recognition based on still images, there is no temporary information available, and therefore the traditional spatio-temporal features can no longer be applied.

In traditional video-based action recognition, low-level features, such as Spatial Point of Interest (STIP), extracted from space-time volume can be used directly for recognizing the actions, but they are not available in still images.

In this type of recognition, generally, low-level features extracted directly from the entire image may not work well. Thus, the papers studied rarely use only the whole image or scene for extraction of low-level features to recognize the actions. Since only spatial information is available on disordered background images, researchers have used different high-level cues in static images to characterize actions rather than using low-level features throughout the image. These high-level cues, which include the human body, body parts, objects, human-object (h-o) interaction, and scene (or context), can be characterized by various low-level features. Then, different high-level cues can be combined to recognize actions in still images.

These cues can characterize human actions from different aspects. A summarization of which cues each study selected in this SLR uses is given in Table 3. From the table, one can see that some approaches employed more cues, while some others used less.

Table 3 – High-level cues

Paper	Human body	Body parts	Objects	H-o interaction	Scene
Weilong Yang et al. (2010)	✓	✓	-	-	-
Delaitre et al. (2010)	✓	✓	-	-	✓
Yao et al. (2011)	✓	✓	✓	-	-
Yao and Fei-Fei (2012b)	✓	✓	✓	-	-
Yao and Fei-Fei (2012a)	✓	-	-	-	-
Sharma et al. (2013)	✓	✓	✓	-	✓
Fahad Shahbaz Khan et al. (2013)	-	✓	✓	-	✓
Zhujin Liang et al. (2014)	-	✓	✓	✓	-
Fahad Shahbaz Khan et al. (2015)	✓	-	-	✓	-
Zhichen Zhao et al. (2016a)	✓	✓	✓	✓	-
Zhichen Zhao et al. (2016b)	-	✓	✓	✓	✓
Lavinia et al. (2016)	-	-	-	-	✓
Zhang et al. (2016)	✓	-	-	✓	✓
Fahad Shahbaz Khan et al. (2016)	-	✓	✓	✓	-
Yan et al. (2017b)	-	✓	✓	-	-
Qi et al. (2017)	✓	✓	-	✓	-
Yan et al. (2017a)	✓	✓	✓	✓	-
Haisheng Zhu et al. (2018)	✓	-	-	-	-
Fahad Shahbaz Khan et al. (2018)	-	✓	✓	✓	-
Haisheng Zhu et al. (2018)	-	✓	-		-
Lu Liu et al. (2018)	-	-	✓	-	✓
Chan et al. (2019)	✓	-	✓	-	✓
Xin et al. (2019)	-	-	-	✓	✓
Mohammadi et al. (2019)	-	-	-	✓	✓
Lavinia et al. (2020)	-	-	-	✓	✓
Siyal et al. (2020)	-	-	-	✓	✓
Lin et al. (2020)	✓	-	-	-	✓
Yu et al. (2020)	-	✓	-	✓	✓
Yunpeng Zheng et al. (2020)	✓	-	-	✓	✓
Ma and Shuang Liang (2020)	✓	-	-	✓	-
Chakraborty et al. (2021)	✓	-	-	✓	✓
Chapariniya et al. (2022)	-	-	-	-	✓
Hirooka et al. (2022)	-	-	-	-	✓
Xiangtao Zheng et al. (2022)	-	-	-	-	✓
Banerjee et al. (2022)	-	✓	-	✓	✓

In practice, it is easier to use the cue scene, especially using the whole image

for action recognition (i.e., no separation between the foreground and background). In contrast, it is relatively difficult to use human body or object cues since it is still challenging to detect the human body or objects automatically with very high accuracy. That is why some public databases provide manually annotated human body bounding boxes, assuming that they are available for action recognition. Finally, it is even harder to extract the cues from body parts or human-object interactions since it is very challenging to detect human body parts or the interactions with a good performance, although these cues could provide more detailed information for action analysis in still images. For example, the action of eating may involve the hand and the mouth, while taking a photo may involve hand and eye (body parts) and a camera (i.e., interaction with object).

Analyzing Table 3, one can conclude that most of the works, especially recently, combine using the scene and H-o interactions. If we look at works published after 2019, the majority of them use H-o interactions and scene as a high-level cue.

3.1.5.3 Low-level features

In the previous subsection, several high-level cues for action recognition on static images were introduced. These high-level cues are usually represented by different low-level features in previous approaches.

In the early stage, Deformable Part Model Detector (DPM) was commonly used, but with the advance of deep leaning-based appr Typical low-level features are represented by a variety of techniques, such as SIFT, DSIFT, HOG, and CNNs. These feature vectors are often encoded with algorithms such as BOW, FV, VLAD, Spatial Pyramid Matching (SPM), among others. These encodings serve similar purposes: summarizing in a vectorial statistic a number of local feature descriptors. Table 4 introduces the low-level features used in each work.

From Table 4, we can observe that convolutional neural networks features have been gaining researchers' attention in recent years, as in several computer vision research fields, replacing SIFT and HOG, and demonstrating their excellent performance also in still image action recognition when compared to other techniques. Authors generally opt for transfer learning from powerful CNNs that were trained in the ImageNet dataset. On the other, the studies analyzed in this SLR do not use state-of-the-art CNNs, but architectures that were introduced a few years ago.

3.1.5.4 Algorithms performance

To understand the current still image-based action recognition status, Table 5 shows the accuracies obtained in the selected approaches in each respective dataset.

As we can verify, SIAR accuracies are still far from other recognition tasks, which are increasingly closer to 100%, as it will be discussed in the next section.

Table 4 – Feature extraction

Paper	Method
Weilong Yang et al. (2010)	poselets + HOG + SVM
Delaitre et al. (2010)	part-based Latent SVM with multi-scale HOG templates with flexible parts (parts detection) + SIFT + BOW
Yao et al. (2011)	poselets + part detectors (atributes and part detectors) + DSIFT and Locality-constrained linear coding (LLC) (classification)
Yao and Fei-Fei (2012b)	DPM + SIFT + spatial pyramid matching SPM + SVM - Mathematical custom model
Yao and Fei-Fei (2012a)	DSIFT + HOG - 2.5D Graph Representation
Sharma et al. (2013)	Expanded Parts Model - DSIFT + Bag of Features
Fahad Shahbaz Khan et al. (2013)	Fusion of Color descriptors + HOG + SIFT + BOW
Zhujin Liang et al. (2014)	DPM + custom Deep Belief Net (DBN) model
Fahad Shahbaz Khan et al. (2015)	DPM + latent SSVM and A-SSVM (detection) + VGG-19 and VGG-16 and linear SVM (classification)
Zhichen Zhao et al. (2016a)	VGG (author did not inform which) + SVM weights for Multi-Scale feature fusion (fixed-bias SVM problem) + SVM
Zhichen Zhao et al. (2016b)	SVM Weights (detect parts) + VGG-16 and VGG-19 + SVM
Lavinia et al. (2016)	VGG-19 + GoogLeNet + ResNet50 + Random Forest and linear SVM.
Zhang et al. (2016)	VGG-16 + PCA + GMM and Fisher Vectors (action mask)
Fahad Shahbaz Khan et al. (2016)	VGG-16 + FV (scale encoding)+ Gaussian mixture model (GMM vocabulary construction)
Yan et al. (2017b)	Fast R-CNN + PCA + VGG-16 + VLAD + SVM
Qi et al. (2017)	VGG-16 (Joint learning)
Yan et al. (2017a)	Multibranch VGG-16 + Attention (Scene and Region)
Haisheng Zhu et al. (2018)	ResNet-50 - Hierarchical Propagation Network
Fahad Shahbaz Khan et al. (2018)	VGG19 + Fisher vector encoding for scale coding + GMM
Lu Liu et al. (2018)	Inception-ResNet-v2 (Two branches: human localization and action classification)
Chan et al. (2019)	Resnet18 (Feature Fusion, Pose, object and scene) + PCA + SVM
Xin et al. (2019)	Branch-fuse network with SCE-loss (entanglement loss) and Resnet-18 backbone
Mohammadi et al. (2019)	NASNet-Large + InceptionResNetV2 + Xception + InceptionV3 (Ensemble)
Lavinia et al. (2020)	GoogleNet, VGG-19, and ResNet-50, ResNet-101 and ResNet-152
Siyal et al. (2020)	GoogleNet, VGG-16, VGG-19, and ResNet-18 + SVM
Lin et al. (2020)	ResNet50 + CPN + CBAM
Yu et al. (2020)	VGG16, VGG19 and Resnet 50 nonsequential CNN + DELWO Deep ensemble learning based on the weight optimization (DELWO)
Yunpeng Zheng et al. (2020)	ResNet-50 + Spatial Attention based Action Mask Networks (SAAM-Net)
Ma and Shuang Liang (2020)	ResNet + self-attention + Faster R-CNN + Bbox Bounding boxes (bbox)
Chakraborty et al. (2021)	VGG-19/ResNet 152/ DenseNet 161 + Bbox
Chapariniya et al. (2022)	ResNet + Attention module
Hirooka et al. (2022)	Ensemble (InceptionV3,Xception, InceptionResnetV2, EfficientNetB7)
Xiangtao Zheng et al. (2022)	EfficientNet-B3 (scene + spatial attention)
Banerjee et al. (2022)	Ensemble (DenseNet 201 + DenseNet 201 with Spatial Attention module)

3.1.6 Findings on research questions

This section discusses how the data extracted from the reviewed studies addresses the research questions. Throughout the investigation, it was sought to provide a synthesized overview of the literature on the main techniques for still image action recognition, in addition to their respective performances.

1. **RQ1 - What are the main techniques used?**

According to this research, it has been identified that a variety of techniques can be used to achieve effective results in all the steps presented by the flowchart in

Table 5 – Performance

Paper \ Dataset	Willow	PASCAL (year)	Stanford 40	PPMI	Other
Weilong Yang et al. (2010)	-	-	-	-	61.07% * [1]
Delaitre et al. (2010)	-	-	-	-	72.16% and 68.76*% [2]
Yao et al. (2011)	-	65.1% (2010)	45.7%	-	-
Yao and Fei-Fei (2012b)	-	-	-	48%	-
Yao and Fei-Fei (2012a)	-	65.12% (2011)	-	43.9%	-
Sharma et al. (2013)	67.6%	-	42.2%	-	-
Fahad Shahbaz Khan et al. (2013)	70.1%	62.4% (2010)	51.9%	-	-
Zhujin Liang et al. (2014) 80.41%	-	-	-	-	-
Fahad Shahbaz Khan et al. (2015)	-	77.0% (2012)	75.4%	-	-
Zhichen Zhao et al. (2016a)	-	90.2% (2012)	78.8%	-	
Zhichen Zhao et al. (2016b)	-	91.6% (2012)	80.6%	-	-
Lavinia et al. (2016)	-	-	81.146%		-
Zhang et al. (2016)	76.96%	83.23% (2012)	82.64%	-	-
Fahad Shahbaz Khan et al. (2016)	92.1%	80.3% (2012)	80.0 %	-	-
Yan et al. (2017b)	-	-	88.5%	81.3%	-
Qi et al. (2017)	-	89.8% (2012)	80.69 %	82.2%	-
Yan et al. (2017a)	-	90.2% (2012)	90.7%	-	-
Haisheng Zhu et al. (2018)	-	-	77.6 %	82.3%	92.5% [3]
Fahad Shahbaz Khan et al. (2018)	92.1%	83.9% (2012)	80 %	82.2 %	-
Haisheng Zhu et al. (2018)	-	86.4% (2012)	-	-	-
Lu Liu et al. (2018)	-	-	94.06%	-	-
Chan et al. (2019)	-	-	-	-	-
Xin et al. (2019)	-	-	87.1%	-	-
Mohammadi et al. (2019)	-	92.1% (2012)	93.17%	-	-
Lavinia et al. (2020)	-	-	84.24%	65.94%	-
Siyal et al. (2020)	-	-	87.22%	-	-
Lin et al. (2020)	-	92.1% (2012)	-	-	-
Yu et al. (2020)	90.19%	-	-	-	96.67% [4]
Yunpeng Zheng et al. (2020)	94.1%	84.8% (2012)	93.0%	-	27.2%
Ma and Shuang Liang (2020)	-	92.8% (2012)	94.6%	-	-
Chakraborty et al. (2021)	-	-	77.2%	85.03%	-
Chapariniya et al. (2022)	-	91.83% (2012)	93.17%	-	
Hirooka et al. (2022)	83.21	-	93.76%	-	-
Xiangtao Zheng et al. (2022)	-	87.8% (2012)	94.6%	-	-
Banerjee et al. (2022)	-	-	87.34%	93.35%	-

[2] Ikizler-Cinbis et al. (2009) dataset
[3] (DELAITRE et al., 2010)
[4] The sports action dataset - (GUPTA et al., 2009)
[5] Actions in still web images: visualization, detection and retrieval - (LI, P. et al., 2011) * Accuracy reported

Figure 13.

Various high-level cues can be used for still image-based action recognition. Since the study on still image-based action recognition is still in the early stage, it is difficult to say which cues are more useful than others. Depending on how to encode the cues and the database used, so far, there is no cue that can always outperform others significantly, although the findings, especially in the recent studies, suggest that combining the scene and H-o interactions may lead to better results.

In the first works on SIAR, to describe the low-level features, HOG and SIFT descriptors, as well as their variations, presented the best results among the consulted papers. With the advent of increasingly powerful computers, techniques based on deep learning have gained more and more space. These approaches have dramatically improved the performance of these state-of-the-art visual recognition systems, highlighting the performance of Convolutional Neural Networks,

which have been achieving state-of-the-art accuracies recently. It is possible to conclude that authors usually apply transfer learning from powerful CNNs that were trained in the ImageNet dataset, but they do not use state-of-the-art CNNs, but architectures that were introduced a few years ago.

The most commonly used classifiers in the state-of-the-art remain the SVMs and their variations, especially the Latent SVM, although there is still intense research on ways to improve the performance of these support vector machines, as well as classification techniques in general. Several papers use CNNs to classify as well as to extract features, while others use CNN just to extract features, combining them with other classifiers, also achieving state-of-the-art results.

It has also proved useful to combine features in multiple scales, using different sources of information within the image. Thus, feature fusion techniques have received wide attention from still image action recognition researchers, as well as related areas in the general recognition field.

2. **RQ2. What is the degree of accuracy compared to other visual recognition tasks?**

Nowadays, researchers on recognition tasks have obtained very precise results, closer to a 100% accuracy rate.

Sorting handwritten digits tasks, for example, reaches 99.79% accuracy in the Modified National Institute of Standards and Technology (MNIST) (WAN et al., 2013) dataset. In the CIFAR-10 set, which has ten classes of images (dogs, cats, etc.), the accuracy rate reaches 96.53% (GRAHAM, 2014). Another well-known dataset is The Street View House Numbers (SVHN) Dataset, which, as the name suggests, deals with street house numbers, reaches 98.31% accuracy (LEE et al., 2016).

Through this study, we verified that the datasets accuracies related to still image action recognition are still far from other recognition tasks. One of the factors pointed out in this research is the difficulty of modeling the scene environment and using this information in the recognition process.

Another factor is the selection of high-level cues since they have different effects and importance for each action recognition. Thus, selecting the most appropriate high-level cues and low-level features to recognize each action is still a challenging problem.

3. **RQ3 - Which high-level cues have the greatest impact on action recognition?**

The initial research in action recognition through static images focused on the complete image to obtain the action classification.

The next step was to find the human in the image using bounding boxes and use this information to extract the features and classify the action, refining this information more and more, such as using poselets and silhouette information. It has also been found that often extracting information from body parts and objects rather than using the body as a whole can lead to better results.

Lately, it has been observed that using object information and its interactions with humans in the image has proved to be another technique that has provided good results.

In this scenario, it is interesting to note the importance of context for the recognition of certain actions. Often, we humans use this contextual information to determine the action on an image. For example, if you see a football field and a ball at the bottom of the image, we often do not need to find humans in the image, determine their silhouette, or search for specific members to tell the action that is being performed.

Note Figure 15. We humans do not need to identify each person in the picture or their respective pose to determine the action being performed. By observing only the background information, the football field, and the general disposition of the people, we can conclude that it is a football match.

Figure 15 – Image representing the action of playing football (REDAÇÃO, 2021).

The major problem today is the difficulty of modeling this contextual information and translating it for use, verifying in the image what is important and what is not for recognition. In addition, unlike video information, a major problem in still image analysis is the lack of motion information. For example, when we see an image with a basketball in the air, like Figure 16, we can easily understand that the ball is in motion. This type of information has not yet been modeled and used in the papers studied and seems a promising path to improve still image action recognition accuracy.

Figure 16 – Image representing the action of throwing a ball (VALDERRAMA, 2021).

3.1.7 Limitations

The main limitations of the review are bias in publication selection and inaccuracy in data extraction. To help ensure that the selection process was unbiased, a research protocol was developed in advance that defined the research questions. Using these questions as a basis, keywords and search terms that would allow us to identify the relevant literature were identified. However, it is important to recognize that software engineering keywords are not standardized and that they can be specific both to the language and the area of interest. Therefore, due to the choice of keywords, there is a risk of relevant studies omission.

To avoid selection bias, each part of the systematic review process was monitored, and in particular the search strategy and citation management procedure, in order to clarify the weaknesses and refine the selection process. Moreover, since the focus of this work was on still image action recognition, it was decided to disregard video analysis.

If the review included such literature, the current study could, in principle, provide more data. In this case, it might have been possible to draw more general conclusions

about action recognition.

To further ensure the unbiased selection of articles, a multi-stage process was used, documenting the inclusion/exclusion reasons at each step, as described in previous sections and also suggested by Barbara Kitchenham (2004).

During the data extraction process, it was found that several articles did not have sufficient details about the methodology or performance and, because of this, the extraction process proved to be complicated. As a consequence, all data from all 31 primary studies were extracted according to a predefined extraction form (Appendix B).

However, the extraction process was often hampered by the way some of the primary studies were reported. Many articles did not have sufficient information for their documentation in a satisfactory way in the extraction form. More specifically, often methods have not been adequately described, and bias and validity issues have not always been addressed. Another limiting factor is the application of certain methods in restricted datasets, which avoids an effective comparison between methods. There is, therefore, the possibility that the extraction process may result in some inaccuracy in the data.

3.2 SLR ON ACTION-BASED IMAGE RETRIEVAL

This second research has the goal to deepen into action-based CBIRs mechanisms that use only still images. It has also been carried out using the guidelines proposed by Barbara Kitchenham (2004) to perform a SLR. The review protocol, identification and selection of primary and secondary studies, data extraction and synthesis, and reporting the results for the reported study are described in the following sections.

The general objective of this review is to answer the following research question:

> RQ. How is the stage of the research on action-based CBIR using only still images?

More specifically, this study addresses the following three issues:

- RQ1.What are the most common approaches for action-based image retrieval that do not use video information?
- RQ2. Which datasets are used for the problem considering only still images and what are the performances?
- RQ3. Which encoding alternatives are used to enhance the algorithm's performance or deal with large datasets?

3.2.1 Data sources and search strategies

Only documents written in English and available online were searched. The search strategy included the same electronic databases used for the first SLR:

- IEEEXplore (www.ieeexplore.ieee.org/Xplore/).
- Scopus (https://www.scopus.com).
- ACM Digital library (www.portal.acm.org/dl.cfm)
- Springer (https://link.springer.com/)
- Biblioteca Nacional de Teses e Dissertações (https://bdtd.ibict.br/vufind/)

Figure 17 presents the systematic review process and the number of documents identified in each stage.

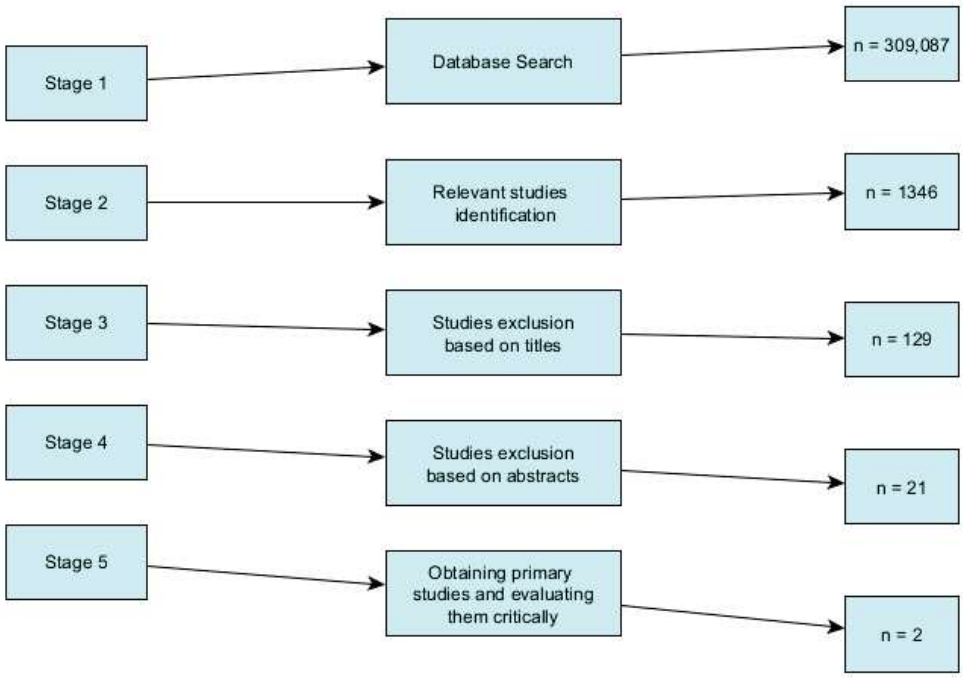


Figure 17 – Stages of the study selection process.

In phase 1, the terms listed in Table 6 were searched.

In this SLR, also editorials, prefaces, article summaries, interviews, news, reviews, correspondences, discussions, comments, readers' letters and tutorial summaries, workshops, panels, and poster sessions were excluded from the search. The research strategy generated a total of 309,087 results that included 1346 non-duplicate citations. Figure 18 presents a visualization of the bibliometric network based on keywords relations considering at least 3 occurrences obtained through the VOSviewer software combining each database.

Table 6 – Search terms used in this review

Term 1	Term 2	Term 3
Action	image	retrieval
Action	CBIR	
Actions	image	retrieval
Actions	CBIR	
Action-based	retrieval	-
Action-based	image	retrieval
Action-based	CBIR	-

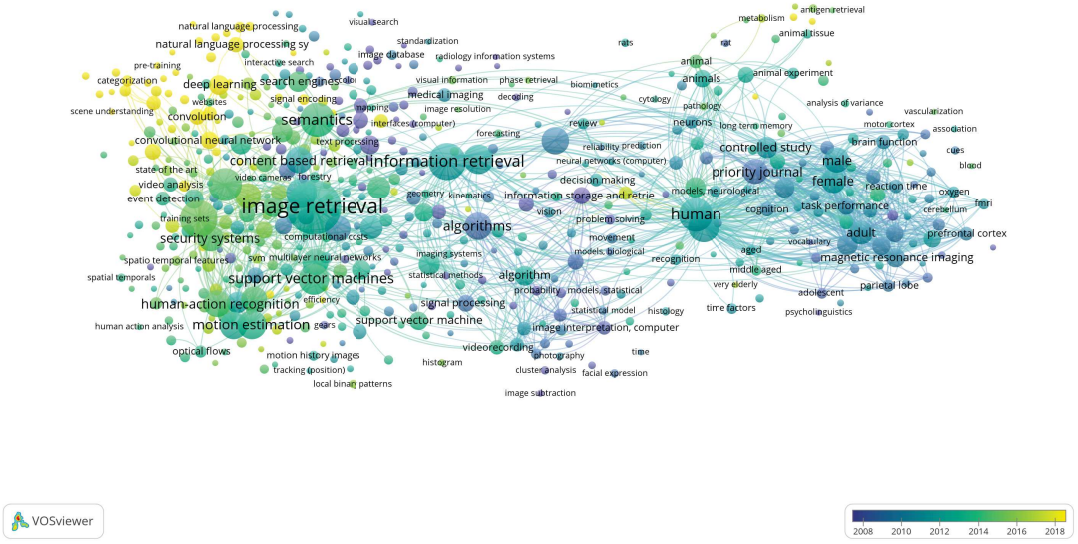


Figure 18 – Bibliometric network based on keywords relations.

It is possible to conclude that research in this topic is closely linked with scene understanding, especially in the last few years, which is one of the main objectives of computer vision.

3.2.2 Managing studies and inclusion decisions

Mendeley was also used to store relevant stage two citation, which were were imported into a Microsoft Excel spreadsheet where the sources of each citation and the subsequent inclusion/exclusion decision were recorded.

In the next stage, the titles of all studies resulting from the previous stage were analyzed to determine their relevance to the systematic review. At this stage, articles with titles that clearly indicated that they were outside the scope of the SLR limit were excluded, leaving 129 studies identified as relevant. However, an article's title does not always represent the document content. During the next stage, each abstract was revised to improve the definition of relevance. Thus, 21 relevant abstracts were found

among the 129 documents.

3.2.3 Final selection

Each of the 21 studies that remained after stage 3 was assessed by the author according to the same three key quality issues that should be considered in the evaluation of the studies identified in the review from three CASP criteria described in the last section (Rigor, Credibility and Relevance); and also the principles of good practices for conducting empirical research in software engineering (KITCHENHAM, B. A. et al., 2002).

Thus, the following selection criteria were used to ensure that the documents address the research topic.

1. Does it use only information provided by still images?
2. Is the purpose of the article clearly stated?
3. Does the document adequately discuss the contextual and technical factors used?
4. Is the proposal tested on public domain datasets?

These four points provided a measure in which the author is confident that a selected paper could be a valuable contribution to fully understand and model the problem. Each of the five criteria was classified on a dichotomous scale (“yes” or “no”), and approved only if it met all the defined criteria.

Two primary studies were selected from the 17 articles in this phase. The quality evaluation was performed based on the selection criteria presented and only works that met the five criteria proposed above were accepted.

The primary and secondary studies selected are enlisted in Appendix B.

3.2.4 Data extraction and synthesis

Following the SLR procedure developed in this thesis, the data were extracted using a predefined data extraction form presented in Appendix C.

The objectives, settings, descriptions of research methods, findings and conclusions, as reported by the primary studies authors were copied into Microsoft Excel.

The data was synthesized by analyzing the discoveries from reported in each of the paper reviewed in this study that could answer one of the research questions, as well as possible clues that would indicate the future of research on this topic.

3.2.5 Results

Two primary studies were selected for this SLR (according to Appendix B). Surprisingly, few studies tackle this specific issue. Some similarities and many differences in each approach were found. They will be presented throughout the next subsections

3.2.5.1 Study overview

Only two studies were selected for this SLR, one from 2011 and other from 2014. The hypothesis is that the decision to focus the research in still image action recognition only in the classification stage, disregarding the detection stage by using bounding boxes at test time is has a major influence in application such as CBIR, since both action recognition and CBIR are a hot topic in computer vision and machine learning, and several studies were found in the researches that address the topic of action-based CBVR Content-Based Video Retrieval (CBVR) (content-based video retrieval).

3.2.5.2 Approaches

Piji Li et al. (2011) built a group of visual discriminative instances for each action class, called “Exemplarlets”, where each exemplarlet is a minimum sub-image (bounding box) to picture the action at global level. They employ Multiple Kernel Learning to learn an optimal combination of histogram intersection kernels, each of which captures a feature channel. In the exemplarlets refining procedure, they manually select about 100 exemplarlets for each action class as positive training examples, and sample about 100 exemplarlets from other action classes as false positives. In the training procedure, the detection is performed with the SVM classifier using the 1-vs-all scheme. For a new image, the approach can detect a “hot-region” using a sliding window detector learned via MKL. The hot-region can imply latent actions in the image. After the hot-region has been detected, an inverted index in the visual search path, called Visual Inverted Index (VII), is built. Finally, fusing the visual search path and the text search path can get the results either relevant to text or visual information. For each action query, they utilize a ranking function to re-organize the result list, treating retrieval task as the re-ranking work, i.e., re-organize the ranking list of text-based search engines.

Elliott et al. (2014), on the other hand, tried to model spatial relationships between image regions using a structured image representation to distinguish between object co-occurrence and interactions. The authors propose to represent the structure of an image using the Visual Dependency Representation, representing an image as a directed acyclic graph over a set of labeled object region annotations, thus, identifying latent representation of the depicted action in the image. The Visual Dependency Grammar defines eight possible spatial relationships between pairs of regions. These relationships in the grammar aimed to provide sufficient coverage of the spatial rela-

tionships required to describe the data and are mathematically defined in terms of pixel overlap, the distance between regions, and the angle between regions. The frame of reference for annotating spatial relationships is the image itself and not the object in the image, and angles and distance measurements are taken or estimated from the centroids of the regions. A trained human annotator creates the Visual Dependency Representation (VDR) of an image in a heavily supervised way. In the first stage, the annotator draws and labels boundaries around the parts of the image they think contribute to defining the action depicted in the image and the context within which the action occurs. In the second stage, the annotator draws labeled directed edges between the annotated regions that capture how the relationships between the image convey the action predict the VDR y of an image over a collection of labeled region annotations x. All of the models used to test the main hypothesis use the cosine similarity function to determine the similarity of the query image to other images in the collection and thus generate a ranked list from the similarity values.

As we can observe, none of these works explore promising deep learning approaches, which have been boosting results over several computer vision tasks, nor employ efficient approaches for encoding/aggregating the images, which may have a great impact on large-scale databases.

3.2.5.3 Datasets

In none of these works, it was found a dataset built exclusively for action-based CBIR. The absence of dataset meant for this purpose made authors pursue two strategies: create a custom dataset or use SIAR datasets. We have to consider the drawback of using a dataset where in many images we already have a centralized action happening, since in practical applications, that is not always the case.

Piji Li et al. (2011) create their own dataset by collecting about 1200 images in total for six action queries: phoning, playing guitar, riding bike, riding horse, running and shooting. Most of the images are collected from Google Image, Bing, Flickr, and others are from PASCAL VOC 2010 and PPMI datasets. On the other hand, Elliott et al. (2014) use PASCAL Visual Object Classification Challenge 2011.

The lack of a standardized dataset for the task also supports the hypothesis for the low number of studies found in this SLR.

3.2.5.4 Algorithms performance

To understand the current action-based still image CBIR status when compared to state-of-the-art CBIR datasets, in this subsection the reported performance of both the studies in this SLR were analyzed.

Piji Li et al. (2011) report the P@100 and mAP for both their approaches, the MKL and KNNK-Nearest Neighbors (KNN) methods. The results are respectively 66%

and 39% in mAP. It is important to highlight that the custom dataset created by the author did not have many actions. Elliott et al. (2014) report the mAP and P@10 of his approaches in the PASCAL Visual Object Classification Challenge 2011. His best results are 51.4% and 45.4% respectively. Similar classes that have less human-object interactions, such as walking, running and jumping, performed very poorly, with both their mAp and p@10 under 22%, which is very far from state-of-the-art approaches for CBIR datasets. Flickr30K 1K (YOUNG et al., 2014) and Oxford 5k (RADENOVIĆ et al., 2018), for example, have several approaches that reach over 80% in mAP or in P@10 even with more classes.

In a scenario where thousands of images are generated each day and the databases' size are getting larger, neither of these two works analyze retrieval time nor time complexity, which are increasingly essential factors since CBIR systems without any care in encoding or aggregation are impractical for these datasets.

3.2.6 Findings on research questions

The conducted SLR allows us to answers the research questions, enlightening the current research status in action-based still image CBIR and showing promise approaches.

1. **RQ1.What are the most common approaches for action-based image retrieval that do not use video information?**

   As few studies have been found on the subject of this SLR, it is not possible to state that there is a dominant technique for action-based CBIR. This is also a reflection of the amount of works in the area and the discourage that using bounding box in test time might bring. When compared to state-of-the-art in CBIR, none of the studies use deep learning approaches, address important issues as time complexity, aggregation, etc.

2. **RQ2. Which datasets are used for the problem considering only still images and whats are the performances?**

   In none of these works it was found a dataset exclusively for action-based CBIR. The absence of dataset meant for this purpose made authors pursue two strategies: create a custom dataset or use standard SIAR datasets. It is important to highlight the drawback of using datasets where in many images we already have a centralized action happening, a scenario very different from many practical applications. The hypothesis is that the lack of a standardized dataset for the task is a hypothesis for the low number of studies found in this SLR.

3. **RQ3. Which encoding alternatives are used to enhance the algorithm performance or deal with large datasets?**

   As stated by Piji Li et al. (2011), the focus of these two studies is tackling the "semantic gap" between action query and visual information of images. Therefore, they treat the retrieval task as trivial re-ranking work. Elliott et al. (2014) encodes the images using VDR, but they do not encode/aggregate those feature vectors using common approaches as LSH or even more efficient approaches as the one proposed by Fan Yang et al. (2019) nor any concern on large-scale datasets. Therefore, none of these studies are concerned with optimizing the features that form the actions codes.

3.2.7 Limitations

This study shares the same limitations as the one developed in the first SLR, highlighting bias in publication selection and inaccuracy in data extraction. The research protocol was developed to minimize this issue and guarantee reproducibility. There is also a risk of relevant studies omission due to the choice of keywords and the possibility that the extraction process may result in some inaccuracy in the data.

3.3 DISCUSSION

Systematic literature reviews on still image action recognition and on action-based image retrieval using only still images were conducted in this chapter. The goal of these reviews was to identify several challenging factors that hinder the static image action representation process and assess the current stage of the research in both fields. Exploring potential strategies to deal with these challenging factors was also another focus of research. The results are presented in two phases: presenting initial quantitative data on the number of papers published each year and the techniques used to extract useful information.

In the second phase of the SLR on SIAR, the data extracted from the primary studies included in these reviews were discussed and analyzed in order to find the research questions' answers. The analysis and interpretation of the data allowed us to draw some general conclusions presented throughout this chapter on the state-of-the-art in still image action recognition and action-based image retrieval using static images.

The results of this review provide information that may be useful for researchers in the large field of computer vision to understand the various challenging factors that may affect the processes of understanding, identifying, recognizing and retrieving images based on action in static images.

Finally, the SLR on action-based still image retrieval helped us discover the current research status in action-based still image CBIR. The hypothesis is that due to the fact that most SIAR approaches focus only on the classification stage, disregarding the detection stage by using bounding boxes at test time, combined with the absence of standard datasets for this end and the difficulty implied in semantic image retrieval, few works tackle practical applications such as the action-based still image CBIR.

The results of this review allow us to clarify the current research status in action-based still image CBIR and to pick some insights, also providing useful information for researchers in the large field of computer vision to understand challenging factors that may affect the processes of retrieving images based on actions when considering only still images.

4 ACT-CBIR: A NOVEL ACTION-BASED STILL IMAGE RETRIEVAL FRAMEWORK

This chapter introduces the Act-CBIR, an action-based CBIR framework that relies only on still images. The two-staged Act-CBIR framework is composed of Dictionary Creation and Image Retrieval. These stages are composed of several modules: an Action Detection module to detect ROIs and extract features from the respective ROI; an Action Encoding and Indexing module to represent each ROI concisely; an Index database to store all ROIs relating them to the base image, and a Similarity Model to retrieve images given a query image. An implementation method using the YOLO v4 architecture as the Action Detector to extract the ROIs from each input image, followed by EfficientNet-B0 CNN backbone retrained to obtain features from each ROI was proposed to experimentally validate the Act-CBIR concept. A binarization of the fully-connected layer (action feature layer) was also proposed, since it uses the efficient hamming distance between codes to decrease computational cost, and large-scale datasets are becoming increasingly popular, and compare it with two other methods for the encoding and indexing pipeline: computing codes directly from the introduced action feature layer and using cosine distance or LSH to retrieve images. Finally, the similarity model retrieves results using an indirect sort using the Quicksort algorithm.

4.1 ACT-CBIR

The Act-CBIR is composed of two main stages: Dictionary Creation and Image Retrieval. The first creates a database with indexes relating the actions found in the image through the action detector, and the latter sorts the image from the database according to the closer index from the query image; and five main modules: Action Detection, Feature Extraction, Action Encoding, Indexing, and Similarity Model. Only still images are required for both training and test time, without the necessity of using video information. Figure 19 summarizes the proposed approach.

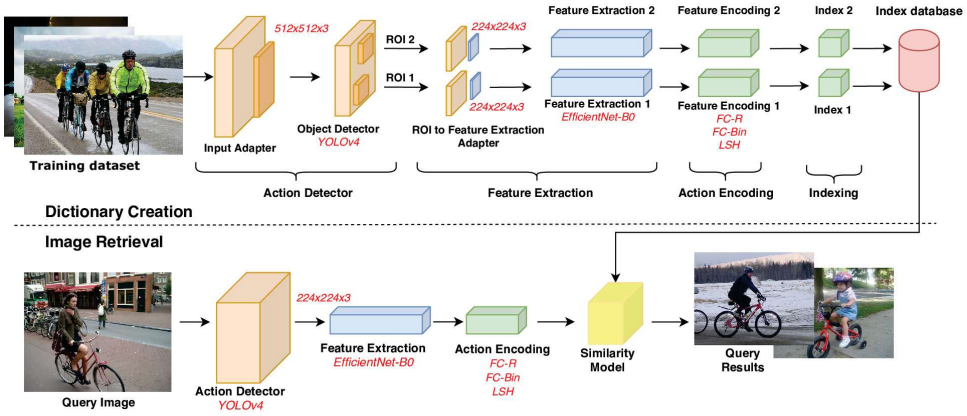


Figure 19 – The general pipeline for the Act-CBIR framework. From the author.

4.1.1 Dictionary Creation

The initial step on building the Act-CBIR is to follow the Dictionary Creation stage to create the index database that stores condensed information of human-object interactions that characterizes an action and relates them to the original image. To create the index database, each training instance must flow through the four modules of the Dictionary Creation pipeline: Action Detection, Feature Extraction, Action Encoding, and Indexing. Further details and explanations of each module are presented in the following subsections.

4.1.1.1 Action Detection

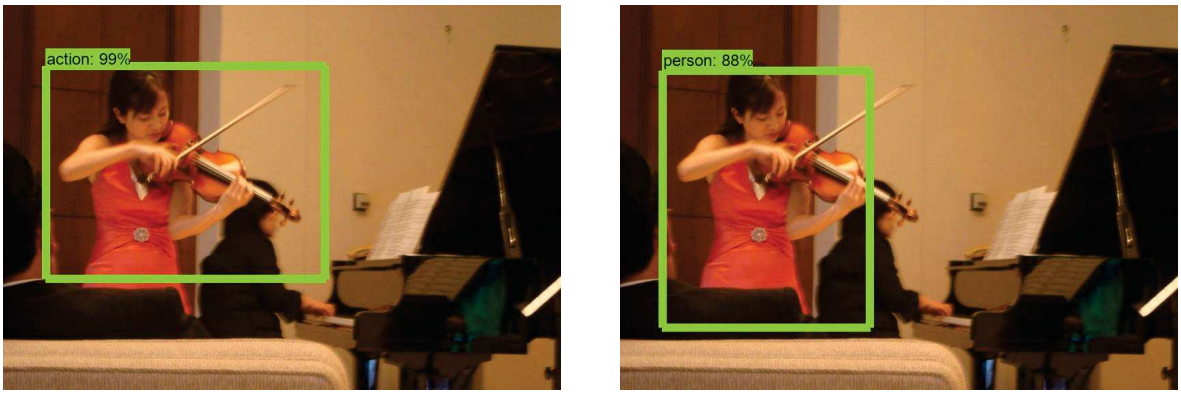
The strong relation between human-object interactions and actions suggests that this information could already be used in the detection stage, providing better results than generic person detectors, and also eliminating the necessity for multiple image patches or the inconvenience of using annotated bounding boxes in test time, which might discourage practical applications.

The first stage of the pipeline for both Dictionary Creation and Image Retrieval is the Action Detection, in which an ROI compressing human-object interactions that characterize actions is extracted. The hypothesis is that this ROI provides better information than ordinary person detectors and also avoids misinformation of cluttered backgrounds presented in full images (the unprocessed image).

Therefore, in the training phase, the images were manually labeled using the LabelImg software[3] with the objective of capturing these human-object interactions that characterize the ROIs. A detection is considered an ROI if it exceeds a minimum threshold when testing, determined experimentally, which means that more than one ROI can be found for one image. Figure 20 presents the result of the action detector compared to a generic person detector.

Object detection has been an intense research topic in recent years, as discussed in Chapter 2. With the help of deep learning approaches achieving superior performance in several recognition tasks, it is possible to observe significant improvements also in object detection, super-passing traditional techniques (ZHAO, Z.-Q. et al., 2019). On the one hand, some applications, such as image retrieval, mobile devices, and many more, often require real-time performance and small memory usage. On the other hand, state-of-the-art methods are optimized for accuracy, which implies that they often rely on model ensembling and multi-crop methods, making them too slow for practical usage, which might impact a query's performance, slowing the process down. Recently, several methods, such as YOLO (REDMON et al., 2016), and SSD (LIU, W. et al., 2016), have been proposed to mitigate running time while maintaining a

[3] https://github.com/tzutalin/labelImg



(a) (b)

Figure 20 – Detection approaches: a) Action Detection; b) Generic Person Detection. The human-object interactions captured by the action detector is highlighted.

decent accuracy. A further discussion on speed/accuracy trade-offs for state-of-the-art detectors can be found at Huang et al. (2017).

Thus, despite the framework should work with several object detectors and produce satisfactory results, it is essential to keep in mind the strengths and weaknesses of each object detection architecture in terms of speed and accuracy and their impact when a query is performed.

4.1.1.2 Feature extraction

Actions are a fine-grained recognition task. After the extraction of these ROIs in an image, features with high discriminative power must extracted due to the nature and similarity of the classes combined with the difficulty of eliminating the semantic gap. In this context, the superior performance of deep learning approaches has resulted in significant improvements in classification and detection tasks. It has been observed that features learned by any layer of deep CNN can serve as generic descriptors for image classification. Furthermore, the discriminative ability of the upper layers is higher than the lower layers because high-level abstractions are modeled in deeper layers (BABENKO et al., 2014). Therefore, it was introduced a fully-connected layer at the end of a CNN architecture, before the classification layer, to compute codes to represent actions in images, which is called FC-R layer, and retrain this CNN in the dataset of interest. The hypothesis is that features from the same class, i.e., same actions, are similar, which would enable an efficient image retrieval.

Similar to the object detector, the feature extraction should work with several architectures that can provide high-level discriminative features. Another critical factor is the feature vector (FC-R) dimension, which could be mitigated by more efficient encoding alternatives, especially for large-scale datasets.

4.1.1.3 Action Encoding

When a query is performed, a similarity ranking is created based on these image representations to filter the images with similar appearances. The upper layers of state-of-the-art CNNs are typically high-dimensional vectors to comprise the information perceived by the network. Therefore, it is imperative to study alternatives to use these features to perform the image retrieval task in this section.

These learned features from the FC-R layer can be directly used to perform image retrieval by comparing the distance (d_c) from the query feature vector (a) to each feature vector (b) learned by the dictionary in the training phase by cosine distance, as shown in Equation 4:

$$d_c(a, b) = \frac{\langle a, b \rangle}{\|(a\| \, \|b\|} \tag{4}$$

However, computing the cosine distance between N-dimensional feature vectors has the time complexity of $O(N^2)$. Thus, it is an expensive operation, which weakens this approach on large-scale collections of images.

Therefore, approaches to reduce the time complexity were explored to make the image retrieval approach feasible in large collections. LSH (INDYK; MOTWANI, 1998) is one of the most popular feature compression methods and is widely used in the field of information retrieval, and computer vision (LIU, R. et al., 2018; GUO, J.; LI, J., 2015). LSH is a set of functions to hash data points into buckets where data points near each other are located in the same buckets with high probability. The basic idea is to generate a k-size hash. Firstly, LSH generates k random hyperplanes in the embedding dimension. Then LSH checks if a particular embedding is above or below the hyperplane and assigns one or zero, repeating for each K hyperplane to obtain the hash value.

When a query using LSH is performed with the hash value of the query image, LSH compares against every item in the buckets.

Due to the limitation of the cosine distance and LSH methods, this thesis proposes an alternative to computing codes to represent actions in images by modifying the FC-R by using the tanh as the activation function. This layer, then, forwards its outputs to a Softmax classifier, which generates probabilities for each class in the training phase. After the training phase, the Softmax layer is removed, and the output of the FC-R is used to represent actions as binary codes after the application of the following threshold function in its activations:

$$\text{FC-Bin}_i = \begin{pmatrix} 1, & \text{Activation}_i \geq 0 \\ 0, & \text{Otherwise} \end{pmatrix} \tag{5}$$

Where *FC-Bin* represents the binarized feature vector for each ROI, $FC\text{-}Bin_i$ its ith bit, and $Activation_i$ represents each ith activation from the FC-R layer. Finally, when

these compact binary codes are obtained, it is possible to use of the hamming distance for comparing the distance (d_H) from the query feature vector (a) to each feature vector (b), which has a time complexity of O(N) and is guided by the Equation 6.

$$d_H(a, b) = a \oplus b \tag{6}$$

4.1.1.4 Indexing

The indexing module organizes the processed information from the previous modules and creates a file that contains the encoded features (*feature*) for each ROI and the corresponding image file path (*file_path*), formally defined as *IndexV=[feature;file_path]*. After processing all images from the training dataset, all *IndexVs* are stacked and stored in a *.h5* file, creating the Index database. Finally, the indexes stored in the database are used by the Similarity Model module in the Image Retrieval stage to get the query results.

4.1.2 Image Retrieval

As depicted in Figure 19, given a query image, the Image Retrieval module consists of five steps: action detection, feature extraction, action encoding, similarity measurement, and query results. The action detection, feature extraction, and encoding steps must follow the processes defined in the Dictionary Creation pipeline, i.e., must use the models and methods. An ROI is a detection that exceeds the same threshold defined in the Dictionary Creation phase, which implies that for one image, more than one search can be done according to the number of ROIs encountered. The encoded features from each ROI are then compared to the features indexed from the dictionary, respecting the encoding choice with the respective similarity measurement. Finally, an indirect sort is performed using the quicksort algorithm in indexes from the dictionary, sorting the images according to the smallest distance.

4.2 IMPLEMENTATION DETAILS

To validate the Act-CBIR framework concept, the Action Detection uses YOLOv4 and for Feature Extraction, a modified EfficientNetB0 (TAN, Mingxing; LE, 2019) as highlighted in red in Figure 19.

In many computer vision applications, higher-capacity networks often lead to superior performance. However, they are often more resource-consuming, which makes it challenging to find models with the right quality-compute trade-off for deployment on edge devices with limited inference budgets. Recently, there has been a growing interest in leveraging Neural Architecture Search (NAS) methods have demonstrated

superior ability in finding models that are not only accurate but also efficient on a specific hardware platform.

During the training of the CNN backbone, the inputs are 224 $\times$ 224 RBG images obtained from the action detector, which are resized without cropping to 224 $\times$ 224 when necessary. No preprocessing is carried out in both cases except for normalization. Before the classification/softmax layer, a fully connected layer of 128 neurons was introduced for all experiments, i.e., FC-R and FC-Bin both have 128 neurons. The size of these layers was defined after experiments and was set the same for both layers for comparison.

Unbalanced data is a frequent problem in datasets, because if we develop a model without considering this disproportionality in the data, the model will face the accuracy paradox, in which the algorithm parameters will not differentiate the minority class of the other categories, believing that they are adding results due to the apparent high accuracy. In order to solve this issue, strategies such as undersampling, which consists of reducing the number of observations of the majority classes to reduce the difference between the categories, and oversampling, which consists of synthetically creating new observations of the minority class, with the aim of matching the proportion of the categories. In this thesis, it was decided to use oversampling in classes with fewer samples by data augmentation to reduce possible imbalance effects.

Data augmentation strategies have also proved to increase model performance by increasing the amount and diversity of data by augmenting it. Data augmentation strategies include flipping, scaling, and changing brightness were applied to adapt the networks for the most diverse inputs. Therefore, each network was initialized with ImageNet weights and fine-tuned it for 1000 steps with RMSprop. Batch normalization was used on convolutional layers, reducing the learning rate on plateaus (factor = 0.1, patience = 10), early stopping (patience = 30), and also a grid search was performed to select the optimal initial learning rate (initial learning rate of the objection detection CNN was 0.001 and 0.005 for the feature CNN), batch size (32 for both CNNs), and dropout rate of 0.4 on the fully connected layer.

Although the model size has increased as a result of the use of the EfficientNet-B0 architecture and the introduction of fully-connected layers (FC-R or FC-Bin), their memory requirements are still lower than other architectures commonly used for recognition, such as the frequently used VGG (SIMONYAN; ZISSERMAN, 2014) and ResNet (SZEGEDY et al., 2017) networks, an essential requirement for the application of this work, which is therefore prioritized. For inference (Dictionary Creation and Image Retrieval), the classification layer is removed and the activations of the last fully-connected layer (FC-R or FC-Bin) are extracted, which is combined with each encoding alternative.

The compressed information composed of each ROI encoded features and the file path of the corresponding training image, learned in the training phase, are stored

in a file forming the database that can be accessed in the query phase. In the query phase, the database images are sorted based on similarity using the corresponding distance in each encoding alternative.

4.3 EXPERIMENTAL RESULTS

This section demonstrates the effectiveness of the proposed approaches. It starts by introducing the datasets followed by the experimental results on public datasets to verify the efficiency of the CBIR both at retrieval time per query and precision.

4.3.1 Datasets

Given the absence of specific datasets for action-based image retrieval using only still images, the experiments were performed in action recognition datasets that contain only still images for both training and testing: the PPMI (YAO; FEI-FEI, 2010), and the PASCAL VOC Action 2012 dataset (EVERINGHAM, M. et al., 2012). Both of these datasets are widely used in recent works and important publications in the area of still image action recognition, as presented in Chapter 3. Each dataset is divided, in a stratified fashion, into train and test, as shown in Table 7. 20% of the train images is taken for validation when the dataset does not provide a validation split.

Table 7 – Datasets

Dataset	Train	Validation	Test
PPMI	891	222	1118
PASCAL VOC Action 2012	2296	2292	2037

PPMI, introduced by (YAO et al., 2011), contains images of humans interacting with twelve different musical instruments: bassoon, cello, clarinet, erhu, flute, horn, guitar, harp, recorder, saxophone, trumpet, and violin, as shown in Figure 21.

It is imperative to highlight that the images may contain multiple people performing different actions, while the image is classified as a single action.

Another dataset used to conduct the experiments is the PASCAL VOC Action 2012 dataset. It consists of 10 different actions: “jumping”, “phoning”, “playing instrument”, “reading”, “riding bike”, “riding horse”, “running”, “taking photo”, “using computer” and “walking”. Evaluation results were obtained in the test images through the “boxless action classification” downloaded from the website. This dataset also contains multiple people performing different actions in one image, causing an image to have different classifications. Figure 22 displays some examples of actions presented in this dataset.

As explained in the subsection 4.2, oversampling by data augmentation strategies was used to reduce class imbalance issues.

Figure 21 – PPMI: Images of people interacting with 12 different musical instruments (YAO et al., 2011)

Figure 22 – The ten class PASCAL VOC 2012 Action dataset. This dataset provides image annotations, which was not used to conduct the experiments

Given the lack of datasets for the specific purpose of CBIR for actions in static images, the choice of these datasets is also due to the fact that they are substantially larger than other datasets commonly used in the still image action recognition task, such as dataset Willow-Actions dataset (DELAITRE et al., 2010) (seven action, 968 images) and UIUC-Sports dataset (LI, L.-J. et al., 2010) (eight sport event categories, 137 to 250 images per sport).

4.3.2 Evaluation Metrics

The mean Average Precision (mAP) has been typically used for evaluation in image retrieval systems (NOH et al., 2017), and it was computed to evaluate the retrieval quality in both datasets. The *mAP* is defined as the mean of the *AP* considering each query q, i.e., considering all query images (Q) used for testing, as stated in Equation 7.

$$mAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{7}$$

The AP@k is another frequently used measure since users often are not interested in the ranking of the whole dataset but the ranking of the top k images. The *AP@k* is defined in Equation 8.

$$AP@k = \frac{\sum_{n}^{k} P@n \times rel(n)}{N_{correct}} \quad (8)$$

Where $N_{correct}$ refers to the total number of ground truth positives, k refers to the total number of interest, $P@n$ is the precision at n, and $rel(n)$ is the relevance function, which is 1 for a relevant retrieval in the rank n of the k images and 0 otherwise. The AP@10 was also adopted in the evaluation protocol, since we are often interested only in the first results of the query. In order to further analyze the performance of the method, extending the AP@10 analysis per each class, considering the three approaches and using cosine distance as the baseline.

4.3.3 Results

This section presents the results of the image retrieval experiment to determine the effectiveness of the introduced framework. In the experiments, a query image from the test collection is used to rank the images from the database, where the goal is to construct a ranking where the top images with the same action as the query image from images in the training set.

Tables 8, 11, 12, and 15 present the results on the PPMI and PASCAL VOC 2012 action datasets, respectively. Meanwhile, Tables 9 and 13 present the retrieval time per query (in milliseconds) for each encoding alternative in each dataset. The results are reported using the action detector, the full image, and a generic person detector. All the images from the test set were used as query images[4].

Table 8 – Results on the PPMI dataset

Model	mAP (%)	AP@10 (%)
Full Image - Cosine	55.19	66.52
Full Image - LSH	54.06	65.64
Full Image - Bin	54.17	65.86
Person - Cosine	39.85	45.03
Person - LSH	38.74	43.87
Person - Bin	38.81	43.30
Action - Cosine	**73.36**	**83.91**
Action - LSH	73.11	82.59
Action - Bin	73.17	82.67

The results comparing approaches confirm the hypothesis that person detectors are sub-optimal for action-based image retrieval. Using the Action detector can increase the mAP up to at least 15% from the second-best method (considering the same encoding alternative), using the full image and obtaining better results than the person

[4] The experiments were run on an Intel Core i7-7700HQ with 16GB RAM and a GeForce GTX 1050Ti.

Table 9 – Retrieval Time per Query (ms) for the PPMI dataset

Model	Retrieval time per query (ms)
Cosine	1.67
LSH	1.45
Bin	**1.22**

detector. Using the cosine distance provided the best results among the encoding techniques, although it is the slowest method. Analyzing the retrieval time per query, it is possible to conclude that the binarized feature vector was the fastest method, proving an excellent trade-off between precision and speed. It was more than 35% faster than using cosine distance and almost 20% faster than LSH.

In order to perform a more thorough quantitative analysis, a statistical analysis of the mAP results of each strategy was performed. Based on studies by Steurer et al. (2021), Gong et al. (2020) and Jianlong Zhou et al. (2021), the results of the Table 10 come from the average value and standard deviation of ten runs of testing. For each of the ten experiments, the training and testing sets are randomly selected using a ten-fold cross-validation scheme.

Table 10 – Statistical Analysis of the PPMI results (Mean $\pm$ SD)

Model	mAP (%)	Retrieval time per query (ms)
Full Image - Cosine	54.86 ± 0.49	1.68 ± 0.01
Full Image - LSH	54.15 ± 0.86	1.51 ± 0.02
Full Image - Bin	53.69 ± 1.01	1.24 ± 0.01
Person - Cosine	39.96 ± 0.62	1.64 ± 0.02
Person - LSH	37.89 ± 0.94	1.42 ± 0.02
Person - Bin	38.56 ± 0.65	1.20 ± 0.01
Action - Cosine	73.39 ± 0.86	1.64 ± 0.01
Action - LSH	73.62 ± 0.13	1.44 ± 0.02
Action - Bin	73.58 ± 0.47	1.19 ± 0.01

Therefore, it is possible to conclude that even considering the worst scenario, the action detector obtained that much better results than the best scenario when compared to other strategies. The same applies to the action encoding strategy, where the proposed strategy easily overcomes LSH and the use of the Cosine distance.

The analysis is extended to each class in the dataset. Table 11 presents the results for each class in these three configurations.

The results demonstrate an enormous advantage in favor of the Action detection, which outperforms the second best method for all actions. The person detection

Table 11 – Results by class in the PPMI dataset AP@10(%)

Model	Full Image	Person Detection	Action Detection
Bassoon	78.13	40.92	89.86
Cello	84.32	56.95	86.02
Clarinet	29.21	27.31	72.65
Erhu	83.19	69.15	92.30
Flute	62.58	32.28	80.13
French Horn	76.94	51.60	82.91
Guitar	93.98	73.67	94.08
Harp	89.61	71.98	90.46
Recorder	38.84	16.94	75.34
Saxophone	34.76	53.77	89.56
Trumpet	44.08	21.73	75.24
Violin	78.51	35.49	86.40

approach, on the other hand, proved to be a deficient alternative, especially for actions with a greater degree of human-object interaction and less importance of their pose.

Table 12 presents results on the PASCAL VOC 2012 Action dataset while Table 13 presents the retrieval time per query of each encoding alternative.

Table 12 – Results on the PASCAL VOC 2012 Action dataset

Model	mAP (%)	AP@10 (%)
Full Image - Cosine	61.28	68.77
Full Image - LSH	60.44	68.25
Full Image - Bin	61.12	68.66
Person - Cosine	40.66	44.71
Person - LSH	39.13	44.64
Person - Bin	39.16	44.54
Action - Cosine	**73.56**	**85.03**
Action - LSH	72.88	84.79
Action - Bin	72.04	84.72

Table 13 – Retrieval Time per Query (ms) for the PASCAL VOC 2012 Action Dataset

Model	Retrieval time per query (ms)
Cosine	1.71
LSH	1.52
Bin	**1.24**

In this dataset, the approaches that used the action detector achieved better results since images with multiple actions are common.

Statistical analysis was also extended to this dataset, as shown in the Table 14.

Table 14 – Statistical Analysis of the PASCAL VOC 2012 Action dataset (Mean $\pm$ SD)

Model	mAP (%)	Retrieval time per query (ms)
Full Image - Cosine	60.84 ± 0.84	1.68 ± 0.01
Full Image - LSH	60.39 ± 0.29	1.49 ± 0.00
Full Image - Bin	61.55 ± 1.03	1.26 ± 0.01
Person - Cosine	40.17 ± 0.78	0.72 ± 0.01
Person - LSH	38.79 ± 0.85	1.51 ± 0.01
Person - Bin	39.04 ± 0.99	1.26 ± 0.01
Action - Cosine	73.80 ± 1.01	1.70 ± 0.00
Action - LSH	72.79 ± 0.87	1.50 ± 0.00
Action - Bin	72.16 ± 0.55	1.25 ± 0.02

Here, it is again possible to ratify the advantages of the action detector and the newly introduced strategy for action encoding.

The analyses was also conducted considering their performance by class for those same three configurations, as shown in Table 15.

Table 15 – Results on the PASCAL VOC 2012 Action dataset per class AP@10(%)

Model	Full Image	Person Detection	Action Detection
Jumping	63.44	41.38	86.64
Phoning	56.63	50.04	86.97
Playing Instrument	84.99	62.33	95.61
Reading	75.21	13.96	88.13
Riding Bike	87.66	63.87	91.67
Riding Horse	91.89	49.40	93.79
Running	59.34	40.71	68.28
Taking Photo	58.02	35.21	87.92
Using Computer	76.45	69.28	88.34
Walking	31.90	17.96	69.06

The analysis of the results in this dataset corroborates the aforementioned conclusions. The Action detection approach, again, showed much better performance than the other approaches by a large margin.

The PASCAL VOC dataset presents more images with multiple actions, which highlights the impact of the action detection stage. The PASCAL VOC 2012 dataset results also corroborate the conclusions presented previously, showing the clear advantages and disadvantages of each method. This suggests that the action detection is very effective, and the efficient hamming distance can be very useful for retrieving images in large-scale datasets.

4.4 DISCUSSION

This chapter presented a framework for image retrieval based on action detection and encoding. Unlike most still image action recognition systems that use additional information from bounding boxes at test time, the proposal is based on an action detection and an encoding approach to content-based image search to retrieve images with people performing similar actions. The thesis opted for architectures with a good exchange trade-off between accuracy and inference time to demonstrate its effectiveness. The approach was evaluated on the public PPMI and PASCAL VOC 2012 Action Recognition datasets, demonstrating an efficient performance in terms of speed and mAP. Each encoding approach had its advantages and drawbacks, but the binarization (FC-Bin), a strategy that introduced by this thesis, proved to be an efficient trade-off in speed and precision, especially when extrapolating to large-scale datasets. The main drawback of this framework is the lack of context information, which may impair its effectiveness for some actions.

5 ACT-RETRIEVAL - AN EVOLUTION FROM THE ACT-CBIR

The last chapter presented the Act-CBIR, a framework for still image action-based image retrieval. Despite the advantages described throughout the chapter, the framework completely disregards any additional information beyond the region of interest, hindering even for us humans to describe some actions.

In order to address inserting context information, this chapter presents the Act-Retrieval, a framework for image retrieval based on multiple inputs, action detection, hint-learning, and an attention module. This framework can consider multiple information. The hypothesis is that this can enhance retrieval performance. Some improvements were introduced in the action representation through the feature vector and in post-processing. Experiments using 12 possible configurations of the framework on publicly available still image action recognition datasets were reported, and an extensive quantitative and qualitative analysis of the results was also performed. All the configurations were evaluated using the standard mAP and AP@10 for each dataset in the quantitative analysis. The AP@10 analyzes was extended by class, pointing out the strengths and weaknesses of each of the 12 possible configurations. The qualitative analyses show the effects of using the action detector, the advantages of the attention module, and even the effect of using multiple input images and combining information by investigating the effect of each module on feature maps.

5.1 DESIGN OVERVIEW

This chapter proposes the Act-Retrieval, a framework for action-based image retrieval that only uses still images and that can consider multiple information, enabling the models to add context information. The training phase, where the dictionary is built, consists of 4 modules: input, feature extraction, action representation, and indexing. The image retrieval phase, i.e., when a query is performed, has five modules: input, feature extraction, action representation, similarity model, and post-processing.

Only still images are required for both training and image retrieval, without the necessity of using video information, human annotations, or bounding boxes at retrieval mode. Figure 23 summarizes the framework.

5.1.1 Configurations

Therefore, the Act-Retrieval framework has 12 possible configurations, according to the path adopted in Figure 23[5] as presented in Table 16.

[5] The reasoning behind the naming involves the input possibilities (F for full image and R for ROI), the branch used, and the main branch indicated by the "+" symbol for cases where both branches are used.

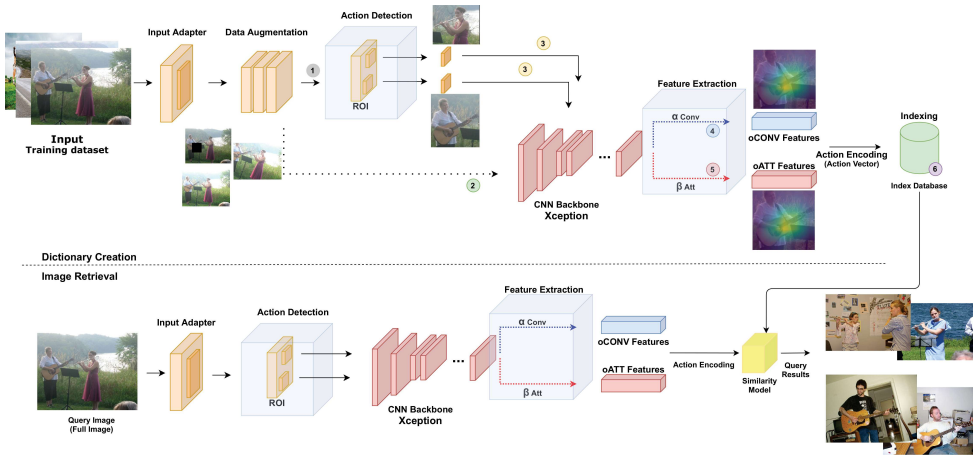


Figure 23 – The two stages of the Act-Retrieval pipeline: Dictionary Creation and Image Retrieval. From the author.

Table 16 – Configurations

Path	Name	Input	CNN Backbone	Branch Features	α	β
1246	F CONV	Full	✓	Convolutional	1.0	0.0
1256	F ATT	Full	✓	Attention	0.0	1.0
124+56	F CONV+	Full	✓	Convolutional	1.0	0.5
1245+6	F ATT+	Full	✓	Attention	0.5	1.0
1346	R CONV	ROI	✓	Convolutional	1.0	0.0
1356	R ATT	ROI	✓	Attention	0.0	1.0
134+56	R CONV+	ROI	✓	Convolutional	1.0	0.5
1345+6	R ATT+	ROI	✓	Attention	0.5	1.0
12346	FR CONV	Full + ROI	✓	Convolutional	1.0	0.0
12356	FR ATT	Full + ROI	✓	Attention	0.0	1.0
1234+56	FR CONV+	Full + ROI	✓	Convolutional	1.0	0.5
12345+6	FR ATT+	Full + ROI	✓	Attention	0.5	1.0

We reinforce that configurations that use the ROI have a pre-step for the action detection.

5.1.2 Input

From the initial image, which is called full image, the framework can use one or two images as inputs: the full image itself and an ROI that should comprise human-object interactions obtained through an object detector built for this purpose. As it was concluded in the last chapter, this ROI provides better information than a generic person or multiple object detectors, avoiding misinformation of cluttered backgrounds presented in full images. Configurations where this ROI can then be combined with the full image, used alone or even not used for image retrieval were tested.

Therefore, the Act-Retrieval has a pre-training phase, where the action detector is trained. The Act-Retrieval pipeline proceeds in the same way as in the previous chapter, the Act-CBIR, where images were manually labeled using the LabelImg software

with the objective of capturing human-object interactions that characterize actions. In the inference stage of the action detector, either in the training phase or in the retrieval mode, a detection is considered an ROI if it exceeds a minimum threshold determined experimentally. Figure 24 presents the action detector's result.

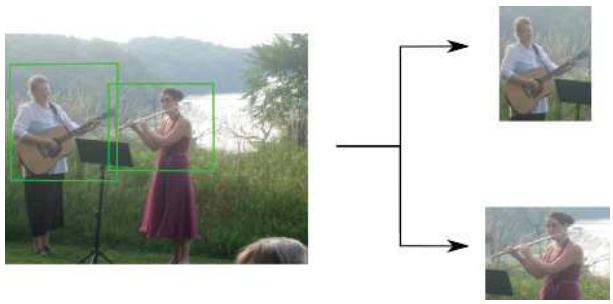


Figure 24 – The action detector selects ROIs that comprises human-object interactions where the object is near a human. From the author.

In configurations that use the action detection module, the number of queries is the same as the number of generated ROIs in retrieval mode.

5.1.3 Feature Extraction

After the input stage, the feature extraction consists of a CNN backbone network and two possible branches: the i) Attention branch and the ii) Convolutional (Conv) branch. The Attention branch is based on the convolutional attention block (WOO et al., 2018), and has a spatial attention module and a channel attention module. The goal of this module is to extract deeper semantic features. The reasoning behind the Conv branch is that the process of identifying an action can be heavily biased by context. Therefore, with slightly "less local" features, the goal is to find the trade-off between both branches, incorporating both local and global features

5.1.3.1 Hint-Learning

Multi-task Learning has been successfully applied in many computer vision and machine learning areas and has also proved to be efficient for still image action recognition (RUDER, 2017; SUDDARTH; KERGOSIEN, 1990; QI et al., 2017). The main idea is to jointly train the CNN for a main and an auxiliary task, which could enhance the representation power and act as a regularizer (QI et al., 2017) since the auxiliary task contributes as a hint to the main task.

Although the attention module helps to find richer feature maps in a region of interest, it often concentrates the feature map in smaller regions, giving less importance to context information that surrounds the object of interest. The strategy is to incorporate context by creating an auxiliary task. To this end, an attention output (oATT, with spatial and color attention modules) and a Conv output (oCONV, a convolutional layer with

64 neurons before the last fully connected layer) are trained together in an end-to-end manner to jointly learn semantic feature representations that better characterize actions. Both of these two branches share all previous layers from the CNN backbone, and the Act-Retrieval uses a loss function ($L(x_i, y_i)$) formally defined in Equation 9 based on the weighted combination (represented by α and β) of each branch for the ith input sample (x_i) and action label (y_i).

$$L(x_i, y_i) = \alpha L_{oCONV}(x_i, y_i) + \beta L_{oATT}(x_i, y_i) \tag{9}$$

In this thesis, the weights α and β assume values of 0, 0.5, and 1, meaning that the Act-Retrieval varies the weight of each branch in the semantic features learning process. The goal is to investigate the impact of fusing the features produced by the attention module and the features found by the convolutional module. In future work, investigations can be conducted to find the optimal values of α and β.

5.1.3.2 Action Encoding

Features from convolutional layers preserve more structural details and provide a high-level semantic representation and are also more robust to image transformations, such as occlusion (BARZ; DENZLER, 2021; SEDDATI et al., 2017). Usually, the robustness of convolutional features is improved after pooling (VACCARO et al., 2020). Therefore, for each branch, the features are max-pooled from the convolutional layer oCONV and oATT, respectively.

As these features are high-dimensional, the feature dimension is reduced by applying the L2-PCA-L2 procedure, where the features are first L2-normalized, followed by PCA dimension reduction, and L2-normalized again, resulting in sum-pooled convolutional features (SPoC) of these descriptors (BABENKO; LEMPITSKY, 2015), named the action vector.

The learned features can be used to perform image retrieval by comparing by comparing the distance (d_c) from the query action vector (a) to each action vector (b) learned by the dictionary in the training phase by cosine distance, as shown in Equation 10:

$$d_c(a, b) = \frac{\langle a, b \rangle}{\|(a\| \, \|b\|} \tag{10}$$

5.1.4 Indexing

The indexing module organizes the processed information from the previous modules and stores the encoded features (*feature*) for each action vector and the corresponding image file path (*file_path*), formally defined as *IndexV* = [*feature*; *file_path*]. After processing all images from the training dataset, all *IndexVs* are stacked and stored

in a *.h5* file, creating the Index database. Finally, the indexes stored in the database are used by the Similarity Model module in the retrieval mode to get the query results. As a result of the action detection, the same image can have several entries in the database derived from the found ROIs, since in an image, several actions can happen at once.

5.1.5 Similarity model

In retrieval mode, after feature extraction and action encoding, an indirect sort using the quicksort algorithm is performed in indexes from the dictionary, ranking the images according to the smallest distance between the action vectors.

5.1.6 Post-processing

Post-processing techniques are often used in image retrieval systems since they have proved to boost results (AZAD; DEEPAK, 2019). In order to enhance the performance of the system, after the ranking creation, a query expansion (EFTHIMIADIS, 1996; CHUM et al., 2007) is performed, where relevant candidates of this first ranking are aggregated into an expanded query, which is then used to re-search images in the database by averaging the top k ranked results and performing another query (CHUM et al., 2007; GORDO et al., 2020). K is empirically set to 10, although further analysis can be done to find the optimal value.

5.2 EXPERIMENTAL PROTOCOL

This section presents the experimental protocol: the implementation details, the datasets used in the experiments (same data augmentation and balancing strategies), and the evaluation criteria for the experiment analysis.

5.2.1 Implementation Details

The framework can use different architectures for object detection and as CNN backbone. Experiments were performed with YOLOv4 (BOCHKOVSKIY et al., 2020) for the action detection, and Xception (CHOLLET, 2017) as CNN backbone as the baseline to prove the effectiveness of the framework due to its good trade-off between accuracy and resource-consuming.

5.2.1.1 Training

In still image actions, we often encounter cluttered images with disordered backgrounds. So, to alleviate this issue, the data augmentation strategies include Cut Mix (YUN et al., 2019), Random Erase (ZHONG et al., 2020), flipping, scaling, and changing brightness to prepare the networks for the most diverse inputs, helping the

semantic generalization. Several regularization techniques were also used, such as Dropout(SRIVASTAVA et al., 2014), label smoothing (SZEGEDY et al., 2016), DropBlock (GHIASI et al., 2018), and confidence penalty (PEREYRA et al., 2017).

The CNNs were initialized with ImageNet weights and fine-tuned for 3000 steps with RMSprop. Batch normalization on convolutional layers was also used, learning rate reducing on plateaus (factor = 0.1, patience = 10), and early stopping (patience = 30). Finally, for hyperparameter tuning, grid search was performed to select the optimal initial learning rate (0.001 for both CNNs), batch size (32 for both CNNs), and dropout rate of 0.4 on the fully connected layer.

The compressed features and the file path associated with each image in the database learned in the training phase are stored in a file forming (.h5) that can be accessed in the query phase. The query image representation is compared with the database in retrieval mode, and the images are sorted based on the similarity measure.

5.2.2 Datasets

The same datasets presented in the previous chapter were used. To recap, Table 17 presents the distribution of each dataset.

Table 17 – Datasets

Dataset	Train	Validation	Test
PPMI	891	222	1118
PASCAL VOC Action 2012	2296	2292	2037

5.2.3 Evaluation Criteria

The Act-Retrieval evaluation followed the same principles presented in the previous chapter, analyzing mAP and Ap@10 in both datasets.

Moreover, the results were compared with reference works and state-of-the-art works on the content-based image retrieval task. Babenko and Lempitsky (2015) is a reference work in CBIR where the authors aggregate local deep features into global descriptors for image retrieval and show that this global descriptor improved the state-of-the-art on four common benchmarks compared to other methods for aggregating convolutional features.

Another very prominent area in CBIR is medical image retrieval. Qayyum et al. (2017) present a reference method to retrieve multimodal medical images for different body organs. As the authors state, this work also tackles the semantic gap in this specific domain. From the presented methodology, an architecture was retrained for the specific purposes of still image action-based image retrieval in each dataset.

The results were also with DeLF (Deep Local Feature) (NOH et al., 2017), (JUN et al., 2019) and (RAMZI et al., 2021). DeLF is an attentive local feature descriptor that yields state-of-the-art performance in several image retrieval datasets, such as Google-Landmarks (NOH et al., 2017), Oxf5k (PHILBIN et al., 2007) and Par6k (PHILBIN et al., 2008) datasets. Jun et al. (2019) proposed a combined descriptor which is generated by concatenating multiple global descriptors in an end-to-end manner to get an ensemble effect, while Ramzi et al. (2021) introduced a loss function in the training set and its averaged batch approximation, and a differentiable approximation of the rank function, which provides an upper bound of the AP loss and ensures robust training achieving state-of-the-art performance in datasets such as iNaturalist Stanford Online Products, CARS196 and CUB200-2011.

Finally, the existing results that in found in the literature were used for comparing. Elliott et al. (2014) obtain their results from the PASCAL Visual Object Classification Challenge 2011 (same actions)[6]. The results were also compared with the method proposed in the previous chapter, the Act-CBIR.

The evaluation was the extended considering AP@10 for each class to analyze the strengths and weaknesses of the approaches using only the F ATT, C ATT, and the combination of information using FR ATT.

A qualitative analysis was also performed, which aims to seek a deeper understanding of the quantitative results, correlating the semantic gap with the performance of the systems for each class and dataset.

5.3 RESULTS

This section presents the results of the image retrieval experiments to determine the effectiveness of the framework. In the experiments, a query image from the test collection is used to rank the images from the database. The goal is to construct a ranking where the images with the same action as the query image from images in the training set. The analysis is divided into quantitative and qualitative analyses.

5.3.1 Quantitative Analysis

This section aims to quantify the performance of the retrieval system in terms of mAP and AP@10 for both datasets. In the experiments, all the 12 possible configurations of the framework on each dataset were tested. Analyses of the performance in each class considering the AP@10 were also carried out, but considering only three approaches: F ATT, C ATT, and FR ATT. The goal was to analyze better the impact of the action detector and the combination of information for each class's nuances.

[6] Xception was selected as backbone for architectures that use CNN backbones for fair comparison.

5.3.1.1 PPMI

Table 18 presents the results on the PPMI action dataset. The results using all of the 12 possible configurations were reported. All the images from the test set are used as query images.

Table 18 – Results of the mAP (%) and AP@10 (%) on the PPMI dataset

Model	mAP(%)	AP@10(%)
FR CONV	71.74	85.26
FR ATT	75.26	89.82
FR CONV+	79.91	88.75
FR ATT+	81.86	**91.71**
F CONV	54.92	68.01
F ATT	56.38	68.08
F CONV+	59.21	71.75
F ATT+	60.37	72.08
R CONV	72.42	86.51
R ATT	74.66	88.99
R CONV+	**83.39**	89.56
R ATT+	83.22	90.49
Act-CBIR	73.36	83.91
Babenko and Lempitsky (2015)	54.62	67.73
Qayyum et al. (2017)	54.88	67.08
Noh et al. (2017)	70.36	78.91
Jun et al. (2019)	72.89	80.43
Ramzi et al. (2021)	58.66	81.94

The results comparing approaches confirm the hypothesis that capturing human-object interactions is imperative for efficient action-based image retrieval. Using the action detector, the Act-Retrieval increases the mAP more than 10% compared to configurations that rely only on the full image. Comparing the multiple-source configurations and configurations that only use the action detector, we cannot reach a clear conclusion of superiority, with a slight advantage for configurations that combine information. Another finding was that, in general, approaches that used the attention module achieved better results, surpassing alternatives that did not use them, which is intuitive when we look at images with actions.

Analogously to the previous chapter, Table 19 presents the statistical analysis of each strategy and comes from the average value and standard deviation of ten runs of testing. Again, the training and testing sets are randomly selected using a ten-fold cross-validation scheme for each of the ten runs.

The results presented in the Table 19 show a wide advantage of the strategies that consider multiple information. Particularly, the strategies that consider the action detector and both branches outperformed the others by some margin.

Table 19 – Statistical Analysis of the PPMI results (Mean $\pm$ SD)

Model	mAP (%)
FR CONV	71.41 $\pm$ 0.66
FR ATT	75.08 $\pm$ 0.39
FR CONV+	80.24 $\pm$ 0.95
FR ATT+	81.88 $\pm$ 0.86
F CONV	54.06 $\pm$ 0.87
F ATT	56.39 $\pm$ 1.02
F CONV+	58.99 $\pm$ 0.73
F ATT+	60.46 $\pm$ 0.91
R CONV	72.53 $\pm$ 0.14
R ATT	74.32 $\pm$ 0.55
R CONV+	81.20 $\pm$ 0.26
R ATT+	82.06 $\pm$ 0.34

In order to better understand the role of the action detector and the combination of information (FR configurations) compared to only using the full image or the ROI, the analysis was extended to each class in the dataset. Table 20 presents the results for each class of these three configurations.

Table 20 – Results by class in the PPMI dataset AP@10(%)

Model	F ATT	R ATT	FR ATT
Bassoon	79.88	92.67	**95.27**
Cello	89.90	**93.82**	92.87
Clarinet	31.06	73.85	**77.05**
Erhu	85.28	97.11	**99.14**
Flute	59.77	82.61	**92.35**
French Horn	78.41	**87.30**	87.22
Guitar	95.79	**97.05**	95.24
Harp	91.20	**97.98**	**97.98**
Recorder	39.33	76.43	**79.88**
Saxophone	36.70	**93.53**	89.67
Trumpet	46.53	77.24	**81.62**
Violin	79.89	**95.66**	92.18

A prevalent problem with the one-image approach (without using the action detector) is that these configurations are able to search for a single action. In this scenario, the use of the ROI found by the action detector proved to be extremely useful, retrieving images with different actions given the multiple queries, yielding better results especially for actions where there is human-object interaction with small objects.

It can be noted a better performance in classes with larger instruments, which could highlight the importance of human-object interactions or even the presence of

these objects in retrieving similar images.

5.3.1.2 Pascal VOC Action

The PASCAL VOC dataset presents more images with multiple actions, which highlights the impact of the action detection stage. The PASCAL VOC 2012 dataset results corroborate the previously presented conclusions, showing each method's clear advantages and disadvantages. Table 21 presents the results on this dataset.

Table 21 – Results on the (%) PASCAL VOC 2012 Action dataset

Model	mAP(%)	AP@10(%)
FR CONV	81.61	87.62
FR ATT	87.56	90.51
FR CONV+	81.12	89.34
FR ATT+	**89.06**	91.34
F CONV	62.41	70.31
F ATT	62.86	72.16
F CONV+	62.97	71.22
F ATT+	63.24	72.47
R CONV	77.62	88.68
R ATT	79.28	88.96
R CONV+	82.10	92.68
R ATT+	84.98	**93.32**
Act-CBIR	73.56	85.03
Elliott et al. (2014)[1]	51.4	45.4[2]
Babenko and Lempitsky (2015)	61.24	70.03
Qayyum et al. (2017)	60.02	68.57
Noh et al. (2017)	76.34	81.70
Jun et al. (2019)	74.45	82.96
Ramzi et al. (2021)	73.13	81.31

[1] PASCAL Visual Object Classification Challenge 2011 (same actions).
[2] P@10

In this dataset, the approaches that used the action detector achieved better results since images with multiple actions are common.

Table 22 presents the statistical analysis of the mAP results in the PASCAL VOC dataset (2012) considering the same method presented earlier.

Again, the results of the statistical analysis confirm the conclusions presented above, as well as the robustness of the network.

The analyses considering their performance by class for those same three configurations was also conducted, as shown in Table 23.

The analysis of the results confirms the analysis obtained in the previous dataset but brings new information: for actions that demand greater context (walking, for exam-

Table 22 – Statistical Analysis of the PASCAL VOC (2012) results (Mean ± SD)

Model	mAP (%)
FR CONV	81.88 ± 0.59
FR ATT	87.49 ± 0.66
FR CONV+	81.74 ± 0.40
FR ATT+	89.08 ± 0.41
F CONV	62.39 ± 0.74
F ATT	62.55 ± 0.77
F CONV+	63.00 ± 0.59
F ATT+	62.89 ± 0.74
R CONV	77.33 ± 0.97
R ATT	79.01 ± 1.14
R CONV+	81.99 ± 0.38
R ATT+	85.12 ± 0.23

Table 23 – Results on the PASCAL VOC 2012 Action dataset per class m@AP10(%)

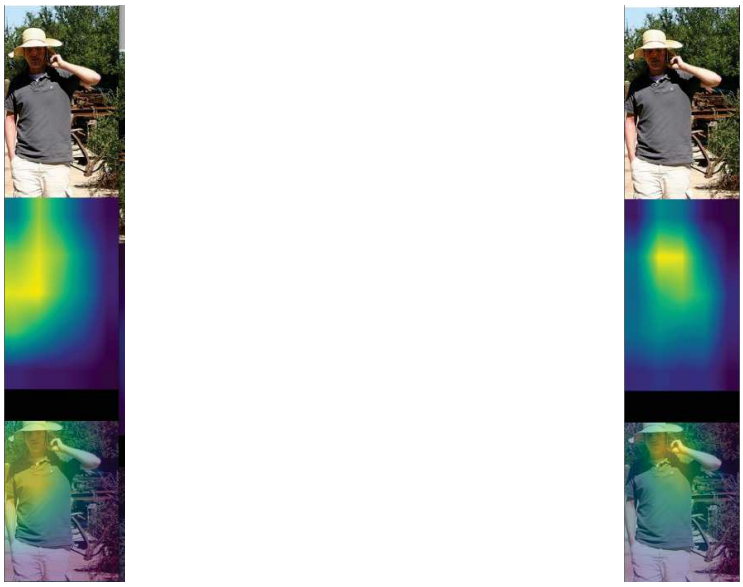
Model	F ATT	R ATT	FR ATT
Jumping	71.74	**88.51**	88.46
Phoning	52.18	89.63	**90.28**
Playing Instrument	86.24	**97.33**	95.25
Reading	87.90	**92.90**	89.06
Riding Bike	88.30	**92.58**	91.39
Riding Horse	91.45	94.35	**96.33**
Running	66.19	71.36	**88.31**
Taking Photo	50.32	**88.24**	87.22
Using Computer	83.87	**91.53**	90.44
Walking	29.83	72.98	**80.97**

ple), using only the ROIs from the action detector proved to be less effective compared to the other classes.

Analyzing the performance of the models by class, we see that the best results occur in classes with more significant interaction between humans and objects, significantly larger objects. In contrast, more abstract classes such as "walking" had the worst results. The hypothesis that these actions require a greater abstraction for interpretation, highlighting the difficulty of dealing with the semantic gap.

5.3.2 Qualitative Analysis

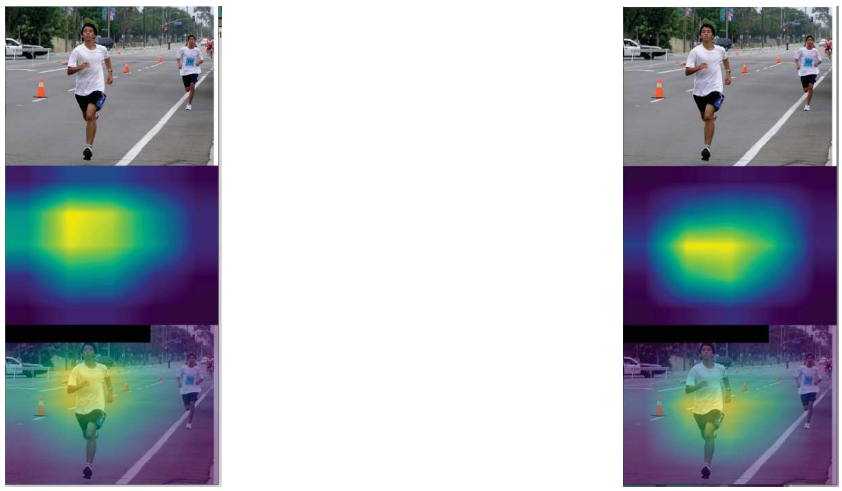
The feature map highlights the regions of images that contribute the most to finding similarities between actions. The first analysis evaluates the impact of the attention module, noting that it contributes significantly more when it comes to actions with greater human-object interaction with small objects, as shown in Figure 25.



(a) (b)

Figure 25 – The attention module displaces the feature map highlighting its importance when dealing with actions with greater interaction with small objects. Figure (a) shows the feature map from CONV and (b) show the displacement produced by the attention module.

Another interesting factor that it was possible to notice when analyzing the feature maps was that the combination of information proved to be useful for classes that need more context information, which is also more evident when we compare for classes that this information is more valuable, such as "running", as shown in Figure 26



(a) (b)

Figure 26 – Using multiple information (full image and ROI) improves performance for actions that need more context information. A displacement of the feature map can be seen from (a) to (b).

The methods that do not make more than one image search are limited to images with only one action, as we see in Figure 27 — in this scenario, using configurations

that have the action detector with multiple ROIs work better.

(a) (b) (c)

Figure 27 – Using the detector with multiple ROIs (Figures b and c, multiple queries) is advantageous for the scenario with many actions on the same image (Figure a).

5.4 DISCUSSION

CBIR has experienced an exponential evolution in the last few years. Nevertheless, semantic image retrieval is still a challenging task. The scenario of retrieving images based on actions in still images is even more precarious. Few works address this issue, and none of these works explore promising deep learning approaches, which have been boosting results over several computer vision tasks. Hence, the Act-Retrieval addresses this important gap in the literature.

This chapter presented a framework for action-based image retrieval that only uses still images. Given the lack of sufficiently challenging datasets for semantic-based image retrieval purposes and even its difficulty, this type of problem is comparatively less studied. The framework works to reduce the semantic gap considering still images and actions. The evaluation shows that methods that produce multiple ROIs outperform methods based on full images. Also, model configurations that use multiple sources of information (multiple inputs or the auxiliary task) in general outperformed others, highlighting the importance of considering multiple sources for semantic-based image retrieval. The results were compared with reference and state-of-the-art works in the image retrieval field, surpassing them by a large margin. Therefore, this work contributes to reduce the semantic gap considering still images and actions, introducing a practical application that can be extendable for several actions.

6 CONCLUSION AND FINAL REMARKS

More and more images are generated daily, a consequence of the popularization of portable devices, which brings the need to find a desired image from a large collection by many professional and personal groups. Unlike meta-data search, which usually does not examine the image's content, relying on textual cues, such as tags and contextual cues that appear near the image, Content-Based Image Retrieval strategies that quantity the image's content have become an ever-increasing need in this scenario.

Humans can provide a rich semantic description of an image with few words easily, unlike machines. This phenomenon is known as the "semantic gap". In this context, human actions provide a natural description of many static images. Human action recognition based on video has been a relatively well-studied research problem in computer vision compared to still image action recognition. Unlike video-based action recognition, where temporal sequences are available and play critical roles, the area of still image-based action recognition requires image content interpretation (ZHANG et al., 2016). The task of image-based action recognition remains a challenging problem due to factors as disordered backgrounds, occlusion, different viewpoints, variations in human posture, changes in lightning, and lack of movements. Despite the seeming advances in CBIR, the gap between the richness of high-level human's perception and low-level machine's descriptions, the semantic gap, has rather become larger than smaller (BARZ; DENZLER, 2021).

In order to fill these gaps, this thesis proposed two action-based image retrieval systems that only use still images and do not need any additional bounding box information when testing, a major drawback for practical applications that arises from the research on still image action recognition.

Initially, two systematic reviews of the literature were carried out to create the basis of this thesis: one in still image action recognition and the second in still image action-based CBIR. The purpose of the SLRs was to examine each field's current research state, summarize and disseminate research findings. With the aid of a robust methodological background, these SLRs made possible to identify research gaps in the literature.

From one of the SLR findings, it was discovered that it is difficult to find publicly available datasets for action-based image retrieval, which is one of the factors the impairs its evolution. The alternative presented by the SLRs was to use publicly available SIAR datasets. Although these image collections usually provide annotations on both train and test time, one could argue that test time annotation makes it unfeasible for applications like CBIR. In order to avoid this issue, it was chosen to create approaches that did not depend on this aspect.

The first framework, Act-CBIR, relies on object detection to capture human-object

interactions, called it the “action detector”. This approach has several advantages since it does not use bounding boxes in the inference stage, i.e., when testing, but it also can capture one of the main high-level cues in still image action recognition, human-object interactions, without using multiple image patches, which would make the system slower. It was also demonstrated that using the entire image (scene) or a generic person detector is suboptimal for the task compared to this action detector that captures h-o interactions. Finally, an strategy for binarizing the feature vector (FC-Bin) aiming to use the efficient hamming distance as a distance measure to compare their similarity was introduced. This strategy proved to be an efficient trade-off in speed and precision, especially when extrapolating large-scale datasets. The effectiveness of this approach was demonstrated in both precision and search time.

Despite the advantages and interesting results of the previous approach, the Act-CBIR completely disregarded any additional information beyond the region of interest captures by the action detector, hindering even for us humans to describe some actions. The second system, Act-Retrieval, is a framework for image retrieval based on multiple inputs, action detection, hint-learning, and an attention module with the goal to overcome this issue. This framework can consider multiple information as input and output, enhancing the retrieval performance, especially for classes heavily dependant on context information, such as "walking". Some improvements in the action representation were introduced, using hint-learning approaches, combining higher-level semantic features and local features to obtain a better action representation.

Therefore, the minor contributions of this thesis are:

- A systematic literature review on still image action recognition;
- A systematic literature review on action-based image retrieval systems that only use still images.

The major contributions of this work are summarized as follows:

- The semantic gap in image retrieval was tackled by proposing novel action-based image retrieval frameworks without exploiting any additional human-made information when testing;
- It was showed that the additional input of human manually made bounding boxes when testing is not necessary and can be replaced by subsystems that capture human-object interactions;
- This thesis demonstrated that the common used generic person detectors are suboptimal for the task of action-based image retrieval;

- The thesis introduced an effective in terms of speed and computational cost encoding method by binarizing the feature vector and using the hamming distance for ranking;
- A framework that can combine multiple sources of information and hint-learning approaches to enhance the CBIR performance especially for actions that heavily depend on context interpretation.
- The effect of multiple information as input was further investigated, using the full image and or an ROI that comprises only the human-object interactions.

6.1 TECHNICAL PRODUCTION

This work has produced as a direct outcome the following technical articles in conferences, symposium, and journals in the area of interest of this thesis:

1. REEBERG DE MELLO, ALEXANDRE ; Stemmer, Marcelo Ricardo ; OLIVEIRA BARBOSA, FLÁVIO GABRIEL . Support vector candidates selection via Delaunay graph and convex-hull for large and high-dimensional datasets. PATTERN RECOGNITION LETTERS, v. 116, p. 43-49, 2018.
2. BARBOSA, F. ; STEMMER, M. R. . ACTION RECOGNITION IN STILL IMAGES BASED ON R-FCN DETECTOR. In: Simpósio Brasileiro de Automação Inteligente (SBAI 2019), 2019, Ouro Preto. Anais do SBAI 2019, 2019. v. 1. p. 1-8.
3. SALAZAR, A. D. ; BARBOSA, F. ; STEMMER, M. R. . Sistema automático para reconhecimento de placas veiculares: comparação entre SVM-HOG e Deep learning. In: Simpósio Brasileiro de Automação Inteligente (SBAI 2019), 2019, Ouro Preto. Anais do SBAI 2019, 2019. v. 1. p. 1-8.
4. BARBOSA, F. G. O. ; REEBERG DE MELLO, ALEXANDRE; STEMMER, M. R. Act-Retrieval: A Framework for Action-Based Image Retrieval. MULTIMEDIA TOOLS AND APPLICATIONS, 2023.
5. BARBOSA, F. G. O. ; REEBERG DE MELLO, ALEXANDRE; STEMMER, M. R. .Act-CBIR: An Action-Based Still Image Retrieval Method. INTERNATIONAL JOURNAL OF PATTERN RECOGNITION AND ARTIFICIAL INTELLIGENCE, 2021. (Submitted. Under review).
6. BARBOSA, F. G. O.; REEBERG DE MELLO, ALEXANDRE; STEMMER, M. R. Act-CBIR: Detecting and Retrieving actions in Still Images. JOURNAL OF VISUAL COMMUNICATION AND IMAGE REPRESENTATION, 2023. (Submitted. Under review).

7. BARBOSA, F. G. O. ; REEBERG DE MELLO, ALEXANDRE; STEMMER, M. R. Act-Retrieval: A Framework for Action-Based Image Retrieval. MULTIMEDIA TOOLS AND APPLICATIONS, 2023. (Accepted).

Below, the relation of each paper with this thesis is detailed:

1. In this paper, strategies for high-dimensional search are explored. These discussed strategies, as well as each respective computational cost, were the basis for the introduced mechanism for action encoding that seeks to reduce computational complexity and search time, described in detail in chapter 4.
2. Prelude to Chapter 4. In this paper, it was demonstrated that using full-images or person detectors are suboptimal for still image action recognition. This paper also introduces the action detector.
3. In this paper, it was applied the concepts of Convolutional Neural Networks, transfer learning, and Convolutional Object Detectors acquired throughout this thesis and described in Chapters 2, 4 e 5, deepening in the aforementioned knowledge.
4. The paper describes the Act-Retrieval, described in Chapter 5.
5. The first version of the Act-CBIR, still no search and analysis time optimization and no comparing considering person detector/full image.
6. The paper describes the Act-CBIR, the second action-based image retrieval built in this thesis and described in detail in Chapter 4.

6.2 LIMITATIONS

The approaches presented in this work have the following drawbacks concerning the strategies used to solve each one of the research objectives:

- Due to the non-existence of a standard dataset for action-based image retrieval systems that only use still images, establishing a baseline for comparison is challenging, since CBIR datasets can point out which images are more similar;
- Both frameworks are heavily supervised, relying on extensive manual annotations for the training phase of the action detector;
- The binarization strategy is designed especially for large-scale datasets, but the additional step of binarizing the FC layer may bring a overhead to small datasets;
- Due to poor word choice both in the search and in the title of articles by the authors, in both SLR relevant works could be omitted. There is also the possibility that the extraction process may result in some inaccuracy in the data, since some

papers did not provide sufficient information that were determined by the review protocol.

6.3 FUTURE WORKS

In line with the main drawbacks of the proposed approaches, there are possible lines of research that look promising for improving them, for example:

- Investigate the impact of several other object detection and classification architectures on the frameworks;
- Create a dataset for action-based image retrieval;
- Find the optimum values of α and β (Chapter 5);
- Study the impact of other loss functions, such as pairwise and triplet ranking losses on the retrieval performance;
- Exploit strategies based on one-shot or few-shot learning;
- Explore quantization techniques;
- Examine strategies for maximizing the relevant information into feature space in a compact way;
- Include state-of-the-art concepts of CBIR mechanisms such as similarity reasoning (DIAO et al., 2021) and other promising attention mechanisms, such as the introduced by the works of Shazeer et al. (2020) or Fanyi Wang et al. (2021);
- Study ways to extract semantic features without the aid of multiple inputs.

6.4 ACKNOWLEDGEMENT

This work was partially supported by the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

ALZU'BI, Ahmad; AMIRA, Abbes; RAMZAN, Naeem. Semantic content-based image retrieval: A comprehensive study. **Journal of Visual Communication and Image Representation**, Elsevier, v. 32, p. 20–54, 2015.

AZAD, Hiteshwar Kumar; DEEPAK, Akshay. Query expansion techniques for information retrieval: a survey. **Information Processing & Management**, Elsevier, v. 56, n. 5, p. 1698–1735, 2019.

BABENKO, Artem; LEMPITSKY, Victor. Aggregating local deep features for image retrieval. In: PROCEEDINGS of the IEEE International Conference on Computer Vision. Parque Araucano, Las Condes, Chile: Institute of Electrical and Electronics Engineers, 2015. P. 1269–1277.

BABENKO, Artem; SLESAREV, Anton; CHIGORIN, Alexandr; LEMPITSKY, Victor. Neural Codes for Image Retrieval. In: SPRINGER. EUROPEAN Conference on Computer Vision. Zurich, Switzerland: Springer, 2014. P. 584–599.

BANERJEE, Avinandan; ROY, Sayantan; KUNDU, Rohit; SINGH, Pawan Kumar; BHATEJA, Vikrant; SARKAR, Ram. An ensemble approach for still image-based human action recognition. **Neural Computing and Applications**, Springer, v. 34, n. 21, p. 19269–19282, 2022.

BARBEE, Anne. **Esportes alternativos conquistam cada vez mais espaço no país**. 2018. Available from: `https://www.portalr3.com.br/2018/04/esportes-alternativos-conquistam-cada-vez-mais-espaco-no-pais/`. Visited on: 4 Apr. 2018.

BARZ, Björn; DENZLER, Joachim. Content-based Image Retrieval and the Semantic Gap in the Deep Learning Era. In: SPRINGER. INTERNATIONAL Conference on Pattern Recognition. Online Streaming: Springer, 2021. P. 245–260.

BAY, Herbert; TUYTELAARS, Tinne; VAN GOOL, Luc. Surf: Speeded up robust features. In: SPRINGER. EUROPEAN Conference on Computer Vision. Graz, Austria: Springer, 2006. P. 404–417.

BELONGIE, Serge; MALIK, Jitendra; PUZICHA, Jan. Shape context: A new descriptor for shape matching and object recognition. In: ADVANCES in neural information

processing systems. Vancouver, British Columbia, Canada: The MIT Press, 2001. P. 831–837.

BOCHKOVSKIY, Alexey; WANG, Chien-Yao; LIAO, Hong-Yuan Mark. YOLOv4: Optimal Speed and Accuracy of Object Detection. **arXiv preprint arXiv:2004.10934**, 2020.

CAI, Deng; GU, Xiuye; WANG, Chaoqi. A revisit on deep hashings for large-scale content based image retrieval. **arXiv preprint arXiv:1711.06016**, 2017.

CAO, Jiuwen; ZHANG, Kai; LUO, Minxia; YIN, Chun; LAI, Xiaoping. Extreme learning machine and adaptive sparse representation for image classification. **Neural networks**, Elsevier, v. 81, p. 91–102, 2016.

CHAKRABORTY, Saikat; MONDAL, Riktim; SINGH, Pawan Kumar; SARKAR, Ram; BHATTACHARJEE, Debotosh. Transfer learning with fine tuning for human action recognition from still images. **Multimedia Tools and Applications**, Springer, p. 1–32, 2021.

CHAN, Abdul Sattar; SALEEM, Kashif; BHUTTO, Zuhaibuddin; MEMON, Mudasar Latif; HUSSAIN, Murtaza; SHAIKH, Saleem Ahmed; SIYAL, Ahsan Raza. Feature Fusion Based Human Action Recognition in Still Images. **International Journal of Computer Science and Network Security**, v. 19, n. 11, p. 151–155, 2019.

CHAPARINIYA, Masoumeh; BARAZANDE, Sara Vesali; ASHRAFI, Seyed Sajad; SHOKOUHI, Shahriar B. Attention Transfer in Self-Regulated Networks for Recognizing Human Actions from Still Images. In: IEEE. 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE). Mashhad, Iran: Institute of Electrical and Electronics Engineers, 2022. P. 036–041.

CHEN, Wei; LIU, Yu; WANG, Weiping; BAKKER, Erwin; GEORGIOU, Theodoros; FIEGUTH, Paul; LIU, Li; LEW, Michael S. Deep image retrieval: A survey. **arXiv preprint arXiv:2101.11282**, 2021.

CHOLLET, François. Xception: Deep learning with depthwise separable convolutions. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: Institute of Electrical and Electronics Engineers, 2017. P. 1251–1258.

CHUM, Ondrej; PHILBIN, James; SIVIC, Josef; ISARD, Michael; ZISSERMAN, Andrew. Total recall: Automatic query expansion with a generative feature model for object retrieval. In: IEEE. 2007 IEEE 11th International Conference on Computer Vision. Rio de Janeiro, Brazil: Institute of Electrical and Electronics Engineers, 2007. P. 1–8.

DAI, Jifeng; LI, Yi; HE, Kaiming; SUN, Jian. R-fcn: Object detection via region-based fully convolutional networks. In: ADVANCES in neural information processing systems. Barcelona, Spain: The MIT Press, 2016. P. 379–387.

DALAL, Navneet; TRIGGS, Bill. Histograms of oriented gradients for human detection. In: IEEE. COMPUTER Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. San Diego, California: Institute of Electrical and Electronics Engineers, 2005. P. 886–893.

DELAITRE, Vincent; LAPTEV, Ivan; SIVIC, Josef. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: BMVC 2010-21st British Machine Vision Conference. Aberystwyth, Wales, United Kingdom: British Machine Vision Association, 2010.

DIAO, Haiwen; ZHANG, Ying; MA, Lin; LU, Huchuan. Similarity Reasoning and Filtration for Image-Text Matching. **arXiv preprint arXiv:2101.01368**, 2021.

DONAHUE, Jeff; JIA, Yangqing; VINYALS, Oriol; HOFFMAN, Judy; ZHANG, Ning; TZENG, Eric; DARRELL, Trevor. Decaf: A deep convolutional activation feature for generic visual recognition. In: INTERNATIONAL conference on machine learning. Beijing, China: JMLR.org, 2014. P. 647–655.

DUBEY, Shiv Ram. A decade survey of content based image retrieval using deep learning. **IEEE Transactions on Circuits and Systems for Video Technology**, Institute of Electrical and Electronics Engineers, 2021.

DYBÅ, Tore; DINGSØYR, Torgeir. Empirical studies of agile software development: A systematic review. **Information and software technology**, Elsevier, v. 50, n. 9, p. 833–859, 2008.

EFTHIMIADIS, Efthimis N. Query Expansion. **Annual review of information science and technology (ARIST)**, ERIC, v. 31, p. 121–87, 1996.

ELLIOTT, Desmond; LAVRENKO, Victor; KELLER, Frank. Query-by-example image retrieval using visual dependency representations. In: PROCEEDINGS of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014. P. 109–120.

EVERINGHAM, M.; VAN GOOL, L.; WILLIAMS, C. K. I.; WINN, J.; ZISSERMAN, A. **The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results**. [S.l.: s.n.], 2012.

EVERINGHAM, Mark; ESLAMI, SM Ali; VAN GOOL, Luc; WILLIAMS, Christopher KI; WINN, John; ZISSERMAN, Andrew. The pascal visual object classes challenge: A retrospective. **International Journal of Computer Vision**, Springer, v. 111, n. 1, p. 98–136, 2015.

GHIASI, Golnaz; LIN, Tsung-Yi; LE, Quoc V. Dropblock: A regularization method for convolutional networks. **arXiv preprint arXiv:1810.12890**, 2018.

GHIMIRE, Deepak; KIL, Dayoung; KIM, Seong-heum. A Survey on Efficient Convolutional Neural Networks and Hardware Acceleration. **Electronics**, MDPI, v. 11, n. 6, p. 945, 2022.

GIRSHICK, Ross. Fast r-cnn. In: PROCEEDINGS of the IEEE international conference on computer vision. Santiago, Chile: Institute of Electrical and Electronics Engineers, 2015. P. 1440–1448.

GIRSHICK, Ross; DONAHUE, Jeff; DARRELL, Trevor; MALIK, Jitendra. Rich feature hierarchies for accurate object detection and semantic segmentation. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. Columbus, OH, USA: Institute of Electrical and Electronics Engineers, 2014. P. 580–587.

GONG, Zhiqiang; ZHONG, Ping; HU, Weidong. Statistical loss and analysis for deep learning in hyperspectral image classification. **IEEE Transactions on Neural Networks and Learning Systems**, Institute of Electrical and Electronics Engineers, v. 32, n. 1, p. 322–333, 2020.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron; BENGIO, Yoshua. **Deep learning**. [S.l.]: MIT press Cambridge, 2016. v. 1.

GORDO, Albert; RADENOVIC, Filip; BERG, Tamara. Attention-based query expansion learning. In: SPRINGER. EUROPEAN Conference on Computer Vision. Glasgow, UK: Springer, 2020. P. 172–188.

GRAHAM, Benjamin. Fractional max-pooling. **arXiv preprint arXiv:1412.6071**, 2014.

GUO, Guodong; LAI, Alice. A survey on still image based human action recognition. **Pattern Recognition**, Elsevier, v. 47, n. 10, p. 3343–3361, 2014.

GUO, Jinma; LI, Jianmin. Cnn Based Hashing for Image Retrieval. **arXiv preprint arXiv:1509.01354**, 2015.

GUPTA, Abhinav; KEMBHAVI, Aniruddha; DAVIS, Larry S. Observing human-object interactions: Using spatial and functional compatibility for recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 31, n. 10, p. 1775–1789, 2009.

HAYKIN, Simon; NETWORK, Neural. A comprehensive foundation. **Neural networks**, v. 2, n. 2004, p. 41, 2004.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Deep residual learning for image recognition. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. San Juan, PR, USA: Institute of Electrical and Electronics Engineers, 2016. P. 770–778.

HE, Xin; ZHAO, Kaiyong; CHU, Xiaowen. AutoML: A Survey of the State-of-the-Art. **Knowledge-Based Systems**, Elsevier, v. 212, p. 106622, 2021.

HINTON, Geoffrey E. Deep belief networks. **Scholarpedia**, v. 4, n. 5, p. 5947, 2009.

HIROOKA, Koki; HASAN, Md Al Mehedi; SHIN, Jungpil; SRIZON, Azmain Yakin. Ensembled transfer learning based multichannel attention networks for human activity recognition in still images. **IEEE Access**, Institute of Electrical and Electronics Engineers, v. 10, p. 47051–47062, 2022.

HU, Benyi; SONG, Ren-Jie; WEI, Xiu-Shen; YAO, Yazhou; HUA, Xian-Sheng; LIU, Yuehu. PyRetri: A PyTorch-based library for unsupervised image retrieval by Deep Convolutional Neural Networks. In: PROCEEDINGS of the 28th ACM International

Conference on Multimedia. Seattle, WA, USA: Association for Computing Machinery, 2020. P. 4461–4464.

HUANG, Jonathan et al. Speed/accuracy trade-offs for modern convolutional object detectors. In: IEEE CVPR. Honolulu, HI, USA: Institute of Electrical and Electronics Engineers, 2017.

IKIZLER, Nazlı; DUYGULU, Pınar. Histogram of oriented rectangles: A new pose descriptor for human action recognition. **Image and Vision Computing**, Elsevier, v. 27, n. 10, p. 1515–1526, 2009.

IKIZLER-CINBIS, Nazli; CINBIS, R Gokberk; SCLAROFF, Stan. Learning actions from the web. In: IEEE. COMPUTER Vision, 2009 IEEE 12th International Conference on. Kyoto, Japan: Institute of Electrical and Electronics Engineers, 2009. P. 995–1002.

INDYK, Piotr; MOTWANI, Rajeev. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. In: ACM. PROCEEDINGS of the thirtieth annual ACM Symposium on Theory of Computing. Dallas, Texas, USA: Association for Computing Machinery, 1998. P. 604–613.

JONES, Simon; SHAO, Ling. Content-based Retrieval of Human Actions from Realistic Video Databases. **Information Sciences**, Elsevier, v. 236, p. 56–65, 2013.

JUN, HeeJae; KO, Byungsoo; KIM, Youngjoon; KIM, Insik; KIM, Jongtack. Combination of multiple global descriptors for image retrieval. **arXiv preprint arXiv:1903.10663**, 2019.

KARPATHY, Andrej. Cs231n: Convolutional neural networks for visual recognition. **Online Course**, 2016.

KHAN, Asifullah; SOHAIL, Anabia; ZAHOORA, Umme; QURESHI, Aqsa Saeed. A survey of the recent architectures of deep convolutional neural networks. **Artificial Intelligence Review**, Springer, v. 53, n. 8, p. 5455–5516, 2020.

KHAN, Fahad Shahbaz; ANWER, Rao Muhammad; VAN DE WEIJER, Joost; BAGDANOV, Andrew D; VANRELL, Maria; LOPEZ, Antonio M. Color attributes for object detection. In: IEEE. COMPUTER Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. Providence, RI, USA: Institute of Electrical and Electronics Engineers, 2012. P. 3306–3313.

KHAN, Fahad Shahbaz; ANWER, Rao Muhammad; WEIJER, Joost van de; BAGDANOV, Andrew D; LOPEZ, Antonio M; FELSBERG, Michael. Coloring action recognition in still images. **International journal of computer vision**, Springer, v. 105, n. 3, p. 205–221, 2013.

KHAN, Fahad Shahbaz; VAN DE WEIJER, Joost; ANWER, Rao Muhammad; BAGDANOV, Andrew D; FELSBERG, Michael; LAAKSONEN, Jorma. Scale coding bag of deep features for human attribute and action recognition. **Machine Vision and Applications**, Springer, v. 29, p. 55–71, 2018.

KHAN, Fahad Shahbaz; WEIJER, Joost van de; ANWER, Rao Muhammad; BAGDANOV, Andrew D; FELSBERG, Michael; LAAKSONEN, Jorma. Scale Coding Bag of Deep Features for Human Attribute and Action Recognition. **arXiv preprint arXiv:1612.04884**, 2016.

KHAN, Fahad Shahbaz; XU, Jiaolong; VAN DE WEIJER, Joost; BAGDANOV, Andrew D; ANWER, Rao Muhammad; LOPEZ, Antonio M. Recognizing actions through action-specific person detection. **IEEE Transactions on Image Processing**, IEEE, v. 24, n. 11, p. 4422–4432, 2015.

KITCHENHAM, Barbara. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, n. 2004, p. 1–26, 2004.

KITCHENHAM, Barbara A; PFLEEGER, Shari Lawrence; PICKARD, Lesley M; JONES, Peter W; HOAGLIN, David C.; EL EMAM, Khaled; ROSENBERG, Jarrett. Preliminary guidelines for empirical research in software engineering. **IEEE Transactions on software engineering**, IEEE, v. 28, n. 8, p. 721–734, 2002.

KRIZHEVSKY, Alex; HINTON, Geoffrey, et al. Learning multiple layers of features from tiny images. Toronto, ON, Canada, 2009.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. In: ADVANCES in neural information processing systems. Lake Tahoe, Nevada, United States: Curran Associates, Incorporated, 2012. P. 1097–1105.

LAVINIA, Yukhe; VO, Holly; VERMA, Abhishek. New colour fusion deep learning model for large-scale action recognition. **International Journal of Computational Vision and Robotics**, Inderscience Publishers (IEL), v. 10, n. 1, p. 41–60, 2020.

LAVINIA, Yukhe; VO, Holly H; VERMA, Abhishek. Fusion Based Deep CNN for Improved Large-Scale Image Action Recognition. In: IEEE. MULTIMEDIA (ISM), 2016 IEEE International Symposium on. San Jose, CA, USA: Institute of Electrical and Electronics Engineers, 2016. P. 609–614.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, Nature Research, v. 521, n. 7553, p. 436–444, 2015.

LEE, Chen-Yu; GALLAGHER, Patrick W; TU, Zhuowen. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In: ARTIFICIAL Intelligence and Statistics. Cadiz, Spain: JMLR.org, 2016. P. 464–472.

LI, Li-Jia; SU, Hao; LI, Fei-Fei; P XING, Eric. Object bank: A high-level image representation for scene classification & semantic feature sparsification. Carnegie Mellon University, 2010.

LI, Piji; MA, Jun; GAO, Shuai. Actions in Still Web Images: Visualization, Detection and Retrieval. In: SPRINGER. INTERNATIONAL Conference on Web-Age Information Management. Wuhan, China: Springer, 2011. P. 302–313.

LI, Zewen; LIU, Fan; YANG, Wenjie; PENG, Shouheng; ZHOU, Jun. A survey of convolutional neural networks: analysis, applications, and prospects. **IEEE Transactions on Neural Networks and Learning Systems**, Institute of Electrical and Electronics Engineers, 2021.

LIANG, Zhujin; WANG, Xiaolong; HUANG, Rui; LIN, Liang. An expressive deep model for human action parsing from a single image. In: IEEE. MULTIMEDIA and Expo (ICME), 2014 IEEE International Conference on. Chengdu, China: Institute of Electrical and Electronics Engineers, 2014. P. 1–6.

LIN, Yixue; CHI, Wanda; SUN, Wenxue; LIU, Shicai; FAN, Di. Human Action Recognition Algorithm Based on Improved ResNet and Skeletal Keypoints in Single Image. **Mathematical Problems in Engineering**, Hindawi, v. 2020, 2020.

LINDSAY, Grace W. Convolutional neural networks as a model of the visual system: Past, present, and future. **Journal of cognitive neuroscience**, MIT Press, v. 33, n. 10, p. 2017–2031, 2021.

LIU, CuiWei; PEI, MingTao; WU, XinXiao; KONG, Yu; JIA, YunDe. Learning a discriminative mid-level feature for action recognition. **Science China Information Sciences**, Springer, v. 57, n. 5, p. 1–13, 2014.

LIU, Lu; TAN, Robby T; YOU, Shaodi. Loss guided activation for action recognition in still images. In: SPRINGER. ASIAN Conference on Computer Vision. Perth, Australia: Springer, 2018. P. 152–167.

LIU, Ruoyu; WEI, Shikui; ZHAO, Yao; YANG, Yi. Indexing of the CNN Features for the Large Scale Image Search. **Multimedia Tools and Applications**, Springer, v. 77, n. 24, p. 32107–32131, 2018.

LIU, Wei; ANGUELOV, Dragomir; ERHAN, Dumitru; SZEGEDY, Christian; REED, Scott; FU, Cheng-Yang; BERG, Alexander C. Ssd: Single shot multibox detector. In: SPRINGER. EUROPEAN conference on computer vision. Amsterdam, The Netherlands: Springer, 2016. P. 21–37.

LOWE, David G. Object recognition from local scale-invariant features. In: IEEE. COMPUTER vision, 1999. The proceedings of the seventh IEEE international conference on. Corfu, Greece: Institute of Electrical and Electronics Engineers, 1999. P. 1150–1157.

MA, Wentao; LIANG, Shuang. Human-Object Relation Network For Action Recognition In Still Images. In: IEEE. 2020 IEEE International Conference on Multimedia and Expo (ICME). London, United Kingdom: Institute of Electrical and Electronics Engineers, 2020. P. 1–6.

MIKOLAJCZYK, Krystian; SCHMID, Cordelia. A performance evaluation of local descriptors. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Institute of Electrical and Electronics Engineers, v. 27, n. 10, p. 1615–1630, 2005.

MOHAMMADI, Sina; MAJELAN, Sina Ghofrani; SHOKOUHI, Shahriar B. Ensembles of Deep Neural Networks for Action Recognition in Still Images. In: IEEE. 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE). Online: Institute of Electrical and Electronics Engineers, 2019. P. 315–318.

NOH, Hyeonwoo; ARAUJO, Andre; SIM, Jack; WEYAND, Tobias; HAN, Bohyung. Large-scale image retrieval with attentive deep local features. In: PROCEEDINGS of

the IEEE international conference on computer vision. Venice, Italy: Institute of Electrical and Electronics Engineers, 2017. P. 3456–3465.

OLIVA, Aude; TORRALBA, Antonio. Building the gist of a scene: The role of global image features in recognition. **Progress in brain research**, Elsevier, v. 155, p. 23–36, 2006.

OQUAB, Maxime; BOTTOU, Leon; LAPTEV, Ivan; SIVIC, Josef. Learning and transferring mid-level image representations using convolutional neural networks. In: IEEE. COMPUTER Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on. Columbus, OH, USA: Institute of Electrical and Electronics Engineers, 2014. P. 1717–1724.

PAN, Sinno Jialin; YANG, Qiang. A survey on transfer learning. **IEEE Transactions on knowledge and data engineering**, IEEE, v. 22, n. 10, p. 1345–1359, 2009.

PEREYRA, Gabriel; TUCKER, George; CHOROWSKI, Jan; KAISER, Łukasz; HINTON, Geoffrey. Regularizing neural networks by penalizing confident output distributions. **arXiv preprint arXiv:1701.06548**, 2017.

PETERSEN, Felix; KUEHNE, Hilde; BORGELT, Christian; DEUSSEN, Oliver. Differentiable top-k classification learning. In: PMLR. INTERNATIONAL Conference on Machine Learning. Baltimore, Maryland, USA: PMLR, 2022. P. 17656–17668.

PHILBIN, James; CHUM, Ondrej; ISARD, Michael; SIVIC, Josef; ZISSERMAN, Andrew. Lost in quantization: Improving particular object retrieval in large scale image databases. In: IEEE. 2008 IEEE conference on computer vision and pattern recognition. Anchorage, AK, USA: Institute of Electrical and Electronics Engineers, 2008. P. 1–8.

PHILBIN, James; CHUM, Ondrej; ISARD, Michael; SIVIC, Josef; ZISSERMAN, Andrew. Object retrieval with large vocabularies and fast spatial matching. In: IEEE. 2007 IEEE conference on computer vision and pattern recognition. Minneapolis, MN, USA: Institute of Electrical and Electronics Engineers, 2007. P. 1–8.

QAYYUM, Adnan; ANWAR, Syed Muhammad; AWAIS, Muhammad; MAJID, Muhammad. Medical image retrieval using deep convolutional neural network. **Neurocomputing**, Elsevier, v. 266, p. 8–20, 2017.

QI, Tangquan; XU, Yong; QUAN, Yuhui; WANG, Yaodong; LING, Haibin. Image-based action recognition using hint-enhanced deep neural networks. **Neurocomputing**, Elsevier, v. 267, p. 475–488, 2017.

RADENOVIĆ, Filip; ISCEN, Ahmet; TOLIAS, Giorgos; AVRITHIS, Yannis; CHUM, Ondřej. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: Institute of Electrical and Electronics Engineers, 2018. P. 5706–5715.

RAMZI, Elias; THOME, Nicolas; RAMBOUR, Clément; AUDEBERT, Nicolas; BITOT, Xavier. Robust and Decomposable Average Precision for Image Retrieval. **Advances in Neural Information Processing Systems**, v. 34, p. 23569–23581, 2021.

REDAÇÃO, O Jogo. **CBIR in the Era of Deep Learning: A Perspective From Feature Representation**. 2021. Available from: `https://www.ojogo.pt/futebol/1a-liga/benfica/noticias/o-lance-do-golo-anulado-a-darwin-no-benfica-sporting-14378929.html`. Visited on: 3 Dec. 2021.

REDMON, Joseph; DIVVALA, Santosh; GIRSHICK, Ross; FARHADI, Ali. You only look once: Unified, real-time object detection. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: Institute of Electrical and Electronics Engineers, 2016. P. 779–788.

REDMON, Joseph; FARHADI, Ali. Yolov3: An incremental improvement. **arXiv preprint arXiv:1804.02767**, 2018.

REN, Shaoqing; HE, Kaiming; GIRSHICK, Ross; SUN, Jian. Faster r-cnn: Towards real-time object detection with region proposal networks. In: ADVANCES in neural information processing systems. Montreal, Quebec, Canada: The MIT Press, 2015. P. 91–99.

RUDER, Sebastian. An overview of multi-task learning in deep neural networks. **arXiv preprint arXiv:1706.05098**, 2017.

SAFAEI, Marjaneh; FOROOSH, Hassan. Still Image Action Recognition by Predicting Spatial-Temporal Pixel Evolution. In: IEEE. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Waikoloa Village, HI, USA: Institute of Electrical and Electronics Engineers, 2019. P. 111–120.

SALEEM, Muhammad Asif; SENAN, Norhalina; WAHID, Fazli; AAMIR, Muhammad; SAMAD, Ali; KHAN, Mukhtaj. Comparative Analysis of Recent Architecture of Convolutional Neural Network. **Mathematical Problems in Engineering**, Hindawi, v. 2022, 2022.

SEDDATI, Omar; DUPONT, Stéphane; MAHMOUDI, Said; PARIAN, Mahnaz. Towards good practices for image retrieval based on CNN features. In: PROCEEDINGS of the IEEE international conference on computer vision workshops. Venice, Italy: Institute of Electrical and Electronics Engineers, 2017. P. 1246–1255.

SHARMA, Gaurav; JURIE, Frédéric; SCHMID, Cordelia. Discriminative spatial saliency for image classification. In: IEEE. COMPUTER Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. Providence, RI, USA: Institute of Electrical and Electronics Engineers, 2012. P. 3506–3513.

SHARMA, Gaurav; JURIE, Frédéric; SCHMID, Cordelia. Expanded parts model for human attribute and action recognition in still images. In: PROCEEDINGS of the IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: Institute of Electrical and Electronics Engineers, 2013. P. 652–659.

SHAZEER, Noam; LAN, Zhenzhong; CHENG, Youlong; DING, Nan; HOU, Le. Talking-heads attention. **arXiv preprint arXiv:2003.02436**, 2020.

SIMONYAN, Karen; ZISSERMAN, Andrew. Very Deep Convolutional Networks for Large-scale Image Recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SIYAL, Ahsan Raza; BHUTTO, Zuhaibuddin; SHAH, Syed Muhammad Shehram; IQBAL, Azhar; MEHMOOD, Faraz; HUSSAIN, Ayaz; AHMED, Saleem. Still Image-Based Human Activity Recognition with Deep Representations and Residual Learning. **International Journal of Advanced Computer Science and Applications (IJACSA)**, v. 11, n. 5, p. 471–477, 2020.

SRIVASTAVA, Nitish; HINTON, Geoffrey; KRIZHEVSKY, Alex; SUTSKEVER, Ilya; SALAKHUTDINOV, Ruslan. Dropout: a simple way to prevent neural networks from overfitting. **The journal of machine learning research**, JMLR. org, v. 15, n. 1, p. 1929–1958, 2014.

STEURER, Miriam; HILL, Robert J; PFEIFER, Norbert. Metrics for evaluating the performance of machine learning based automated valuation models. **Journal of Property Research**, Taylor & Francis, v. 38, n. 2, p. 99–129, 2021.

SUDDARTH, Steven C; KERGOSIEN, YL. Rule-injection hints as a means of improving network performance and learning time. In: SPRINGER. EUROPEAN Association for Signal Processing Workshop. Sesimbra, Portugal: Springer, 1990. P. 120–129.

SZEGEDY, Christian; IOFFE, Sergey; VANHOUCKE, Vincent; ALEMI, Alexander A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: THE Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, California USA: AAAI Press, 2017. P. 12.

SZEGEDY, Christian; VANHOUCKE, Vincent; IOFFE, Sergey; SHLENS, Jon; WOJNA, Zbigniew. Rethinking the inception architecture for computer vision. In: PROCEEDINGS of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA: Institute of Electrical and Electronics Engineers, 2016. P. 2818–2826.

SZELISKI, Richard. **Computer vision: algorithms and applications**. [S.l.]: Springer Science & Business Media, 2010.

TAN, Mingxing; CHEN, Bo; PANG, Ruoming; VASUDEVAN, Vijay; SANDLER, Mark; HOWARD, Andrew; LE, Quoc V. Mnasnet: Platform-aware neural architecture search for mobile. In: PROCEEDINGS of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: Institute of Electrical and Electronics Engineers, 2019. P. 2820–2828.

TAN, Mingxing; LE, Quoc. Efficientnet: Rethinking model scaling for convolutional neural networks. In: PMLR. INTERNATIONAL conference on machine learning. Long Beach Convention & Entertainment Center, Long Beach, Califórnia, USA: [s.n.], 2019. P. 6105–6114.

THURAU, Christian; HLAVÁC, Václav. Pose primitive based human action recognition in videos or still images. In: IEEE. COMPUTER Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. Long Anchorage, AK, USA: Institute of Electrical and Electronics Engineers, 2008. P. 1–8.

TORREY, Lisa; SHAVLIK, Jude. Transfer learning. **Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques**, v. 1, p. 242, 2009.

VACCARO, Federico; BERTINI, Marco; URICCHIO, Tiberio; DEL BIMBO, Alberto. Image retrieval using multi-scale CNN features pooling. In: PROCEEDINGS of the 2020 international conference on multimedia retrieval. Dublin, Ireland: Association for Computing Machinery, 2020. P. 311–315.

VALDERRAMA, Álvaro; CNN. **CBIR in the Era of Deep Learning: A Perspective From Feature Representation**. 2021. Available from: `https://cnnespanol.cnn.com/gallery/la-carrera-de-kobe-bryant-en-fotos/`. Visited on: 26 Jan. 2021.

VAN DE WEIJER, Joost; SCHMID, Cordelia. Coloring local feature extraction. In: SPRINGER. EUROPEAN conference on computer vision. Graz, Austria: Springer, 2006. P. 334–348.

WAN, Li; ZEILER, Matthew; ZHANG, Sixin; CUN, Yann L; FERGUS, Rob. Regularization of neural networks using dropconnect. In: PROCEEDINGS of the 30th international conference on machine learning (ICML-13). Atlanta, GA, USA: JMLR.org, 2013. P. 1058–1066.

WANG, Fanyi; HU, Haotian; SHEN, Cheng. BAM: A Lightweight and Efficient Balanced Attention Mechanism for Single Image Super Resolution. **arXiv preprint arXiv:2104.07566**, 2021.

WANG, Jingdong et al. Deep high-resolution representation learning for visual recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Institute of Electrical and Electronics Engineers, 2020.

WANG, Lei. **CBIR in the Era of Deep Learning: A Perspective From Feature Representation**. 2016. Available from: `https://www.slideshare.net/XiaohuZHU/cbir-in-the-era-of-deep-learning?from_action=save`. Visited on: 15 Aug. 2016.

WOO, Sanghyun; PARK, Jongchan; LEE, Joon-Young; KWEON, In So. Cbam: Convolutional block attention module. In: PROCEEDINGS of the European conference on computer vision (ECCV). Munich, Germany: Springer, 2018. P. 3–19.

WORTSMAN, Mitchell et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: PMLR. INTERNATIONAL Conference on Machine Learning. Baltimore, Maryland, USA: PMLR, 2022. P. 23965–23998.

WU, Xinhui; XIAO, Shuangjiu. Sketch-Based Image Retrieval via Compact Binary Codes Learning. In: SPRINGER. INTERNATIONAL Conference on Neural Information Processing. Siem Reap, Cambodia: Springer, 2018. P. 294–306.

XIE, Michael; JEAN, Neal; BURKE, Marshall; LOBELL, David; ERMON, Stefano. Transfer learning from deep features for remote sensing and poverty mapping. **arXiv preprint arXiv:1510.00098**, 2015.

XIN, Miao; WANG, Shuhang; CHENG, Jian. Entanglement Loss for Context-Based Still Image Action Recognition. In: IEEE. 2019 IEEE International Conference on Multimedia and Expo (ICME). Shanghai, China: Institute of Electrical and Electronics Engineers, 2019. P. 1042–1047.

YAN, Shiyang; SMITH, Jeremy S; LU, Wenjin; ZHANG, Bailing. Multi-branch Attention Networks for Action Recognition in Still Images. **IEEE Transactions on Cognitive and Developmental Systems**, Institute of Electrical and Electronics Engineers, 2017a.

YAN, Shiyang; SMITH, Jeremy S; ZHANG, Bailing. Action Recognition from Still Images Based on Deep VLAD Spatial Pyramids. **Signal Processing: Image Communication**, Elsevier, v. 54, p. 118–129, 2017b.

YANG, Fan; HINAMI, Ryota; MATSUI, Yusuke; LY, Steven; SATOH, Shin'ichi. Efficient image retrieval via decoupling diffusion into online and offline processing. In: 01. PROCEEDINGS of the AAAI Conference on Artificial Intelligence. Honolulu, Hawaii, USA: AAAI Press, 2019. P. 9087–9094.

YANG, Weilong; WANG, Yang; MORI, Greg. Recognizing human actions from still images with latent poses. In: IEEE. COMPUTER Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. San Francisco, CA, USA: IEEE, 2010. P. 2030–2037.

YAO, Bangpeng; FEI-FEI, Li. Action recognition with exemplar based 2.5 d graph matching. **Computer Vision–ECCV 2012**, Springer, p. 173–186, 2012a.

YAO, Bangpeng; FEI-FEI, Li. Grouplet: A structured image representation for recognizing human and object interactions. In: IEEE. COMPUTER Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. San Francisco, CA, USA: IEEE, 2010. P. 9–16.

YAO, Bangpeng; FEI-FEI, Li. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 34, n. 9, p. 1691–1703, 2012b.

YAO, Bangpeng; JIANG, Xiaoye; KHOSLA, Aditya; LIN, Andy Lai; GUIBAS, Leonidas; FEI-FEI, Li. Human action recognition by learning bases of action attributes and parts. In: IEEE. COMPUTER Vision (ICCV), 2011 IEEE International Conference on. Barcelona, Spain: IEEE, 2011. P. 1331–1338.

YOUNG, Peter; LAI, Alice; HODOSH, Micah; HOCKENMAIER, Julia. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. **Transactions of the Association for Computational Linguistics**, MIT Press, v. 2, p. 67–78, 2014.

YU, Xiangchun; ZHANG, Zhe; WU, Lei; PANG, Wei; CHEN, Hechang; YU, Zhezhou; LI, Bin. Deep ensemble learning for human action recognition in still images. **Complexity**, Hindawi, v. 2020, 2020.

YUE, Jun; LI, Zhenbo; LIU, Lu; FU, Zetian. Content-based image retrieval using color and texture fused features. **Mathematical and Computer Modelling**, Elsevier, v. 54, n. 3-4, p. 1121–1127, 2011.

YUN, Sangdoo; HAN, Dongyoon; OH, Seong Joon; CHUN, Sanghyuk; CHOE, Junsuk; YOO, Youngjoon. Cutmix: Regularization strategy to train strong classifiers with localizable features. In: PROCEEDINGS of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. P. 6023–6032.

ZHANG, Yu; CHENG, Li; WU, Jianxin; CAI, Jianfei; DO, Minh N; LU, Jiangbo. Action recognition in still images with minimum annotation efforts. **IEEE Transactions on Image Processing**, IEEE, v. 25, n. 11, p. 5479–5490, 2016.

ZHAO, Zhichen; MA, Huimin; CHEN, Xiaozhi. Multi-scale region candidate combination for action recognition. In: IEEE. IMAGE Processing (ICIP), 2016 IEEE International Conference on. Phoenix, AZ, USA: IEEE, 2016a. P. 3071–3075.

ZHAO, Zhichen; MA, Huimin; CHEN, Xiaozhi. Semantic parts based top-down pyramid for action recognition. **Pattern Recognition Letters**, Elsevier, v. 84, p. 134–141, 2016b.

ZHAO, Zhong-Qiu; ZHENG, Peng; XU, Shou-tao; WU, Xindong. Object Detection with Deep Learning: A Review. **IEEE Transactions on Neural Networks and Learning Systems**, Institute of Electrical and Electronics Engineers, v. 30, n. 11, p. 3212–3232, 2019.

ZHENG, Xiangtao; GONG, Tengfei; LU, Xiaoqiang; LI, Xuelong. Human action recognition by multiple spatial clues network. **Neurocomputing**, Elsevier, v. 483, p. 10–21, 2022.

ZHENG, Yunpeng; ZHENG, Xiangtao; LU, Xiaoqiang; WU, Siyuan. Spatial attention based visual semantic learning for action recognition in still images. **Neurocomputing**, Elsevier, v. 413, p. 383–396, 2020.

ZHONG, Zhun; ZHENG, Liang; KANG, Guoliang; LI, Shaozi; YANG, Yi. Random erasing data augmentation. In: 07. PROCEEDINGS of the AAAI Conference on Artificial Intelligence. New York, NY, USA: AAAI Press, 2020. P. 13001–13008.

ZHOU, Jianlong; GANDOMI, Amir H; CHEN, Fang; HOLZINGER, Andreas. Evaluating the quality of machine learning explanations: A survey on methods and metrics. **Electronics**, MDPI, v. 10, n. 5, p. 593, 2021.

ZHOU, Wengang; LI, Houqiang; TIAN, Qi. Recent advance in content-based image retrieval: A literature survey. **arXiv preprint arXiv:1706.06064**, 2017.

ZHU, Fan; SHAO, Ling; XIE, Jin; FANG, Yi. From handcrafted to learned representations for human action recognition: A survey. **Image and Vision Computing**, Elsevier, v. 55, p. 42–52, 2016.

ZHU, Haisheng; HU, Jian-Fang; ZHENG, Wei-Shi. Learning Hierarchical Context for Action Recognition in Still Images. In: SPRINGER. PACIFIC Rim Conference on Multimedia. Hefei, China: Springer, 2018. P. 67–77.

ZHUO, Li; CHENG, Bo; ZHANG, Jing. A comparative study of dimensionality reduction methods for large-scale image retrieval. **Neurocomputing**, Elsevier, v. 141, p. 202–210, 2014.

Appendix

APPENDIX A – WORKS INCLUDED IN THE SIAR REVIEW

Primary Studies:

[*S*1] YANG, W.; WANG, Y.; MORI, G. Recognizing human actions from still images with latent poses. In: IEEE. **Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on**. Honolulu, Hawaii,USA: AAAI Press, 2010. p. 2030–2037.

[*S*2] DELAITRE, V.; LAPTEV, I.; SIVIC, J. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: **BMVC 2010-21st British Machine Vision Conference**. Aberystwyth, Wales, United Kingdom:British Machine Vision Association, 2010.

[*S*3] YAO, B.; FEI-FEI, L. Action recognition with exemplar based 2.5 d graph matching. **Computer Vision–ECCV 2012**, Springer, p. 173–186, 2012.

[*S*4] YAO, B.; FEI-FEI, L. Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 34, n. 9, p. 1691–1703, 2012.

[*S*5] YAO, B. et al. Human action recognition by learning bases of action attributes and parts. In: IEEE. **Computer Vision (ICCV), 2011 IEEE International Conference on**. Barcelona, Spain: IEEE, 2011. p.1331–1338.

[*S*6] SHARMA, G.; JURIE, F.; SCHMID, C. Expanded parts model for human attribute and action recognition in still images. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition**. . Portland, OR, USA: IEEE, 2013. p. 652–659.

[*S*7] KHAN, F. S. et al. Coloring action recognition in still images. **International journal of computer vision**, Springer, v. 105, n. 3, p. 205–221, 2013.

[*S*8] LIANG, Z. et al. An expressive deep model for human action parsing from a single image. In: IEEE. **Multimedia and Expo (ICME), 2014 IEEE International Conference on**. Chengdu, China: IEEE, 2014. p. 1–6.

[*S*9] KHAN, F. S. et al. Recognizing actions through action-specific person detection. **IEEE transactions on image processing**, IEEE, v. 24, n. 11, p. 4422–4432, 2015.

[*S*10] ZHANG, J.; HAN, Y.; JIANG, J. Tucker decomposition-based tensor learning for human action recognition. **Multimedia Systems**, Springer, v. 22, n. 3, p. 343–353, 2016.

[*S*11] ZHAO, Z.; MA, H.; CHEN, X. Multi-scale region candidate combination for action recognition. In: IEEE. **Image Processing (ICIP), 2016 IEEE International Conference on**. Phoenix, AZ, USA: IEEE, 2016. p.3071–3075.

[*S*12] ZHAO, Z.; MA, H.; CHEN, X. Semantic parts based top-down pyramid for action recognition. **Pattern Recognition Letters**, Elsevier, v. 84, p. 134–141, 2016

[*S*13] LAVINIA, Y.; VO, H. H.; VERMA, A. Fusion based deep cnn for improved large-scale image action recognition. In: IEEE. **Multimedia (ISM), 2016 IEEE International Symposium on**. San Jose, CA, USA: IEEE, 2016. p. 609–614.

[*S*14] ZHANG, Y. et al. Action recognition in still images with minimum annotation efforts. **IEEE Transactions on Image Processing**, IEEE, v. 25, n. 11, p. 5479–5490, 2016.

[*S*15] KHAN, F. S. et al. Scale coding bag of deep features for human attribute and action recognition. **arXiv preprint arXiv**:1612.04884, 2016.

[*S*16] YAN, S.; SMITH, J. S.; LU, W.; ZHANG, B. Multi-branch attention networks for action recognition in still images. **IEEE Transactions on Cognitive and Developmental Systems**, IEEE, 2017.

[*S*17] YAN, S.; SMITH, J. S.; ZHANG, B. Action recognition from still images based on deep vlad spatial pyramids. **Signal Processing: Image Communication**, Elsevier, v. 54, p. 118–129, 2017.

[*S*18] QI, T. et al. Image-based action recognition using hint-enhanced deep neural networks. **Neurocomputing**, Elsevier, v. 267, p. 475–488, 2017.

[*S*19] ZHU, H.; HU, J.-F.; ZHENG, W.-S. Learning hierarchical context for action recognition in still images. In: SPRINGER. **Pacific Rim Conference on Multimedia.** Hefei, China: Springer, 2018. p. 67–77.

[*S*20] LI, R.; LIU, Z.; TAN, J. Reassessing hierarchical representation for action recognition in still images. **IEEE Access**, IEEE, 2018.

S[21] KHAN, F. S.; WEIJER, J. van de; ANWER, R. M.; BAGDANOV, A. D.; FELSBERG, M.; LAAKSONEN, J. Scale coding bag of deep features for human attribute and action recognition. **Machine Vision and Applications**, Springer, v. 29, n. 1, p. 55–71, 2018.

S[22] CHAN, Abdul Sattar; SALEEM, Kashif; BHUTTO, Zuhaibuddin; MEMON, Mudasar Latif; HUSSAIN, Murtaza; SHAIKH, Saleem Ahmed; SIYAL, Ahsan Raza. Feature Fusion Based Human Action Recognition in Still Images. **International Journal of Computer Science and Network Security**, v. 19, n. 11, p. 151–155, 2019.

S[23] XIN, Miao; WANG, Shuhang; CHENG, Jian. Entanglement Loss for Context-Based Still Image Action Recognition. In: IEEE. 2019 IEEE International Conference on Multimedia and Expo (ICME). Shanghai, China: IEEE, 2019. P. 1042–1047.

S[24] MOHAMMADI, Sina; MAJELAN, Sina Ghofrani; SHOKOUHI, Shahriar B. Ensembles of Deep Neural Networks for Action Recognition in Still Images. In: IEEE. 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE). Online Conference:IEEE, 2019. P. 315–318.

S[25] LAVINIA, Yukhe; VO, Holly; VERMA, Abhishek. New colour fusion deep learning model for large-scale action recognition. **InternationalInternational Journal of Computational Vision and Robotics**, Inderscience Publishers (IEL), v. 10, n. 1, p. 41–60, 2020.

S[26] SIYAL, Ahsan Raza; BHUTTO, Zuhaibuddin; SHAH, Syed Muhammad Shehram; IQBAL, Azhar; MEHMOOD, Faraz; HUSSAIN, Ayaz; AHMED, Saleem. Still Image-Based Human Activity Recognition with Deep Representations and Residual Learning. **International Journal of Advanced Computer Science and Applications (IJACSA)**, v. 11, n. 5, p. 471–477, 2020.

S[27] LIN, Yixue; CHI, Wanda; SUN, Wenxue; LIU, Shicai; FAN, Di. Human Action Recognition Algorithm Based on Improved ResNet and Skeletal Keypoints in Single Image. **Mathematical Problems in Engineering**, Hindawi, v. 2020, 2020.

S[28] YU, Xiangchun; ZHANG, Zhe; WU, Lei; PANG, Wei; CHEN, Hechang; YU, Zhezhou; LI, Bin. Deep ensemble learning for human action recognition in still images. **Complexity**, Hindawi, v. 2020, 2020.

S[29] ZHENG, Yunpeng; ZHENG, Xiangtao; LU, Xiaoqiang; WU, Siyuan. Spatial attention based visual semantic learning for action recognition in still images. **Neurocomputing**, Elsevier, v. 413, p. 383–396, 2020.

S[30] MA, Wentao; LIANG, Shuang. Human-Object Relation Network For Action Recognition In Still Images. In: IEEE. 2020 IEEE International Conference on Multimedia and Expo (ICME). London, United Kingdom: IEEE, 2020. P. 1–6.

S[31] CHAKRABORTY, Saikat; MONDAL, Riktim; SINGH, Pawan Kumar; SARKAR, Ram; BHATTACHARJEE, Debotosh. Transfer learning with fine tuning for human action recognition from still images. **Multimedia Tools and Applications, Springer**, p. 1–32, 2021.

S[32] CHAPARINIYA, Masoumeh; BARAZANDE, Sara Vesali; ASHRAFI, Seyed Sajad; SHOKOUHI, Shahriar B. Attention Transfer in Self-Regulated Networks for Recognizing Human Actions from Still Images. In: IEEE. 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE). Mashhad, Iran: IEEE, 2022. P. 036–041.

S[33] HIROOKA, Koki; HASAN, Md Al Mehedi; SHIN, Jungpil; SRIZON, Azmain Yakin. Ensembled transfer learning based multichannel attention networks for human activity recognition in still images. IEEE Access, IEEE, v. 10, p. 47051–47062, 2022.

S[34] ZHENG, Xiangtao; GONG, Tengfei; LU, Xiaoqiang; LI, Xuelong. Human action recognition by multiple spatial clues network. Neurocomputing, Elsevier, v. 483, p. 10–21, 2022.

S[35] BANERJEE, Avinandan; ROY, Sayantan; KUNDU, Rohit; SINGH, Pawan Kumar; BHATEJA, Vikrant; SARKAR, Ram. An ensemble approach for still image-based human action recognition. Neural Computing and Applications, Springer, v. 34, n. 21, p. 19269–19282, 2022.

Secondary Studies:

[*S*36] GUO, G.; LAI, A. A survey on still image based human action recognition. **Pattern Recognition**, Elsevier, v. 47, n. 10, p. 3343–3361, 2014.

[*S*37] ZIAEEFARD, M.; BERGEVIN, R. Semantic human activity recognition: a

literature review. **Pattern Recognition**, Elsevier, v. 48, n. 8, p. 2329–2345, 2015.

[*S*38] ZHU, F. et al. From handcrafted to learned representations for human action recognition: A survey. **Image and Vision Computing**, Elsevier, v. 55, p. 42–52, 2016.

APPENDIX B – WORKS INCLUDED IN THE ACTION-BASED IMAGE RETRIEVAL REVIEW

Primary Studies:

[*S*1] LI, Piji; MA, Jun; GAO, Shuai. Actions in Still Web Images: Visualization, Detection and Retrieval. In: SPRINGER. INTERNATIONAL Conference on Web-Age Information Management. Wuhan, China: Springer, 2011. P. 302–313.

[*S*2] ELLIOTT, Desmond; LAVRENKO, Victor; KELLER, Frank. Query-by-example image retrieval using visual dependency representations. In: PROCEEDINGS of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Dublin City University and Association for Computational Linguistics, 2014. P. 109–120.

APPENDIX C – DATA EXTRACTION FORM

Table 24 – Data Extraction Form

Study Description		
1.	Study identifier	Unique identifier for each study
2.	Extraction date	-
3.	Bibliographical reference	Author, year, title, source
4.	Article type	Journal article, conference document, workshop, book section
5.	Study Objectives	What were the objectives?
6.	Description	Description of problem addressed
7.	Research hypothesis	Hypothesis statement, if any
8.	Used Methods	Methods used to solve the problem
9.	Results	Accuracy and their respective databases
10.	Result Analysis	Comparison with state of the art
Study Findings		
11.	Findings and Conclusions	What were the findings and conclusions?
12.	Validity	Limitations, threats to validity
13.	Relevance	Research, practice