UNIVERSIDAD NACIONAL DE COLOMBIA

Human action video retrieval

Fabián Mauricio Páez Rivera

Universidad Nacional de Colombia
Engineering School, Systems and Industrial Engineering Department
Bogotá, Colombia
2015

Human action video retrieval

Fabián Mauricio Páez Rivera

Thesis presented in fulfillment of a partial requirement to apply for the degree of:
Master on Systems Engineering and Computing

Advisor:
Ph.D. Fabio Augusto González Osorio

Research topic:
Machine learning, information retrieval, computer vision
Research group:
MindLAB

Universidad Nacional de Colombia
Engineering School, Systems and Industrial Engineering Department
Bogotá, Colombia
2015

To my parents, who gave me the gift of life, and
to my brother Ruben, whose influence has played
an important role in my life

Acknowledgements

I want to thank to all the people who helped me in any way during my master studies. First of all, I thank my parents and brothers for their support, company and comprehension. I also want to thank Colciencias and the Jovenes Investigadores grant, which allowed me to work full time on my thesis. And last but not least, thanks to Professor Fabio Gonzalez for his teachings, guidance and support and also to all the members of the MindLAB research group with whom I have had the opportunity to work and learn.

Abstract

The problem of efficiently answering a user information need in a video collection related to human actions is addressed in this thesis. The focus is given to the case where the user queries are stated using an example video containing the action of interest. Among the motivations of the work is the growing complexity of available video content in terms of size and content diversity, and also the ubiquity of video content fueled by the widespread use of video cameras. To solve the problem at hand, an information retrieval system is proposed where multiple information modalities are leveraged if available to discover the latent semantics of the video collection. The central component are matrix factorization-based indexes which have been previously used on image retrieval settings. Along the way, different features and encoding methods for the visual information have been evaluated, such as Bag of Features, Fisher Vectors and Improved Trajectory Features. As a result, a system achieving similar performance as Support Vector Machines-based systems has been obtained.

Keywords: latent semantics, information retrieval, multimodal indexing, matrix factorization, video analysis.

Resumen

El problema de responder eficientemente a la necesidad de información de un usuario en una colección de video relacionada con acciones humanas es abordado en esta tesis. El enfoque es dado al caso donde las consultas del usuario son planteadas usando un video de ejemplo conteniendo la acción de interes. Entre las motivaciones del trabajo esta la creciente complejidad del contenido de video disponible en terminos de tamaño y diversidad de contenido, y tambien a la ubicuidad de contenido de video potenciado por la amplia difusión de camaras de video. Para resolver el problema a la mano, se propone un sistema de recuperación de información en donde multiples modalidades de información son aprovechadas si están disponibles, para descubrir la semántica latente de la colección de videos. El componente central son índices basados en factorización de matrices que han sido utilizados previamente en configuraciones de recuperación de imágenes. En el camino, diferentes caracteristicas y métodos de codificacion para la información visual han sido evaluados, tales como Bolsa de caracteristicas, Vectores de Fisher y Caracteristicas de Trayectorias Mejoradas. Como resultado, se ha obtenido un sistema que logra desempeño similar a sistemas basados en Maquinas de Vectores de Soporte.

Palabras clave: semántica latente, recuperación de información, indexación multimodal, factorización de matrices, análisis de video.

Contents

1 Introduction

Information retrieval is defined in [42] as "finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored in computers)", see Figure 1.1 for an illustration. The term unstructured in that definition, contrasts with the common conception of a database which is structured from the relation between data [20], allowing simple information querying. This unstructured data poses the need to find appropriate methods to perform queries in a collection.
This is a problem easily illustrated in the domain of web pages. Their content is unstructured even though there have been efforts to give it structure by means of the semantic web [3]. Despite the efforts, its adoption wasn't́ as expected and methods focused on unstructured collections are still common. In the case of a web page, a user has an information need and expresses it as a textual query presented to the search engine and the relevant results to the query are obtained. The search engine takes the user query and through diverse techniques it finds the relations among the query and the collection, which in this case is made of the web pages crawled and indexed by the search engine [42].
A general workflow of an information retrieval system is made of:

1. Feature extraction: By means of a given algorithm, features are extracted with the aim of obtaining the relevant contents of the collection.

2. Index generation: With the features extracted on the previous step, an index is built for the collection.

3. Query processing: The user query is processed in a similar way as the collection during the feature extraction step. This is made to obtain a common representation where comparisons can be carried.

4. Similarity measure calculation: Some similarity measure is used to determine the degree of match between query and collection items.

The collections need not be of textual content. There are also methods to retrieve on audio, image and video collections that roughly follow the aforementioned steps. As these collections have information contained in different modalities, multimodal retrieval strategies may be applied that simultaneously involve the different sources. This collections present additional challenges to those found on textual retrieval. First of all, there is the problem of presenting the query to the system. Among the alternatives to solve the problem are keyword-based

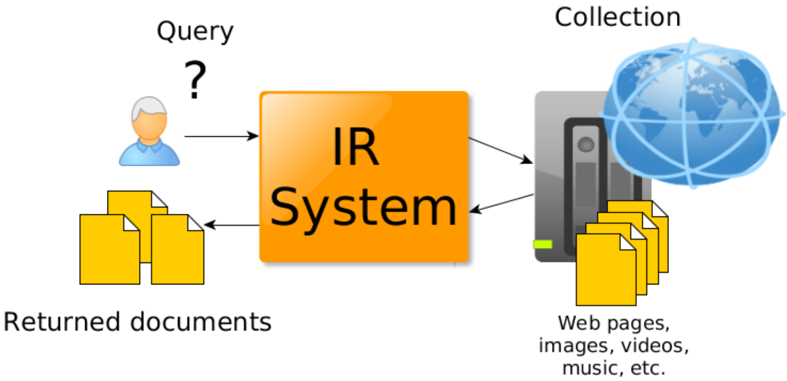


Figure 1.1: Diagram illustrating an information retrieval system. The user has an information need that is expressed to the system as a query. In response the system returns to the user a set of documents from the collection, which the system considers relevant to the query.

queries and query by example [16] which consists on presenting the system an example of an image, video, etc. and based on the example the system finds similar elements. There is also the problem of multimodality and the semantic gap [19]. The semantic gap refers to the difference in meanings between various descriptions of the same object. While one or more words may give a very specific meaning, an image may have multiple interpretations and the problem lies in developing systems that close this gap and find the semantic relationships between the representations of different modalities.

1.1 Problem statement

During the last years electronic devices with growing processing capacity and smaller sizes and prices have been massified, e.g. tablets, smartphones, digital cameras. This trend has allowed every person with a smartphone or digital camera to record, store and publish video content. Along with this phenomena, social networks have been developed which allow to share multimedia content and have received wide acceptance. According to the ”Ericsson Mobility Report” [12], video has the first place on data traffic in mobile networks (31%) for 2012 and its forecasted to grow and reach almost half of the traffic (46%) in year 2018. These two phenomena have raised the need to develop systems that allow to search and explore multimedia collections.

The problem addressed in this thesis is the retrieval of human actions in video collections. The interest in studying video rises initially by the previously mentioned phenomenon. Additionally, emphasis is made on human actions because I consider that as video is produced by humans, they would also be starring most of the available content. Therefore, the actions present in a video would be a good discriminative factor when performing searches.

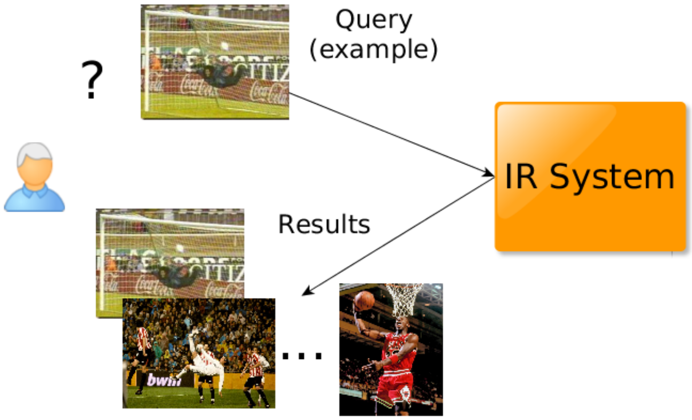


Figure 1.2: In the case of a video collection, the user may query the system using the query-by-example method. The user would give an example video of the relevant content to his needs and the system returns relevant videos. Relevance may be measured by the match in objects, context, motions or a mixture of those.

Human action video retrieval has applications in surveillance, athlete training and performance assessment, traffic analysis, education/training and human behavior analysis in general. The objective is that given a user query by means of keywords or examples, to find in a video collection occurrences of videos containing the action in the user query. This would be useful in the case of surveillance to check who, where and when has performed an specific action. It may also be used in marketing to evaluate the interests of customers of a market and how they interact with products and publicity [54]. To achieve a good performance in a video retrieval system, many aspects of information retrieval and video processing must be taken into account.

First, it is necessary to find ways of representing the multimodal content present in the video so that searching is based on the most relevant features. Additionally, querying methods are required which simplify to the user the process of expressing its information needs, it could be by means of keywords or examples of the searched content. Figure 1.2 presents a diagram of an information retrieval system using the query by example method. Finally, indexing and collection-query comparison strategies must be implemented which are at the same time efficient, robust and tolerant to variability intrinsic to video content such as occlusion, viewpoint and scale variation, etc.

1.2 Objectives

The main objective of the thesis has been to design and implement an strategy for the representation, indexing and retrieval of human actions from video collections. To achieve this objective, the work has been divided in achieving the following specific objectives, which are focused in the main components of a retrieval system:

- Evaluate the different collections of human action videos and the select one for experimentation. The ideal collection would be comparable to the ones indexed by search engines, so the focus for this component is on large scale relating to actions and number of samples.

- Adapt or design methods for video content representation, aimed at capturing properly the information of actions.

- Adapt or design indexing strategies that allow to perform the retrieval efficiently under limited processing and storage capabilities.

- Implement an information retrieval system with the selected representation and indexing methods

- Systematically evaluate the implemented system on an appropriate video collection

1.3 Contributions

The main contribution of the thesis is the design of an information retrieval system using state-of-the-art datasets, features and indexing strategies. As products of the research, the following articles have been submitted to conferences:

- "An Evaluation of NMF Algorithm on Human Action Video Retrieval", accepted and presented at the Simposio de Tratamiento de Señales, Imágenes y Vision Artificial 2013 (STSIVA 2013).

- "Online Multimodal Matrix Factorization for Human Action Video Indexing", accepted and presented at the Content Based Multimedia Indexing 2014 (CBMI 2014).

- "Annotating and Retrieving Human Actions using Matrix Factorization", awaiting for response from the reviewers of the Congreso Iberaoamericano de Reconocimiento de Patrones 2015 (CIARP 2015).

Most of the code used is available in web repositories:

- The code for the method presented in the work submitted to CBMI 2014 is available at
  https://bitbucket.org/fmpaezri/ifomf

- The contributions made to the codebase of [63] which where used for the work submitted to CIARP 2015, are available at
  https://bitbucket.org/fmpaezri/semantic-embedding-methods

- Code for the representation pipeline is available at https://bitbucket.org/fmpaezri/thumos.

1.4 Thesis organization

Chapter 2 gives a brief state of the art relating datasets, features and human action retrieval. The research starts with an evaluation of batch methods based on matrix factorization in Chapter 3. The next step was using an unsupervised online matrix factorization method to speed up training, reduce memory requirements and to be able to process larger datasets, presented in Chapter 4. Finally, in Chapter 5, an online method with an objective function which includes a supervised term is applied for retrieval and its performance in annotation is also evaluated, closing with the main conclusions on Chapter 6.

2 State of the art

Video information retrieval poses the typical problems on the domain of information retrieval. Additional problems are posed due to the high information content, even multimodal e. g. audio, text and image. Within this area is of particular interest the retrieval of human actions, as its applications are diverse such as in surveillance, sports and marketing. Closely related to this task is human action recognition and as they share a common study subject, ideas from this area can be used to solve the retrieval task.

This state of the art presents a brief introduction to the video retrieval problem. The different strategies and features developed so far are reviewed and the most common datasets are enumerated and described.

2.1 Datasets

Research in human action recognition has been made for many years and many approaches have been proposed. Compared to human action retrieval, recognition is a mature field and a lot of publications can be found. Many datasets focused on action recognition have been released but few on retrieval. Therefore recognition datasets are the first candidate to be used in retrieval.

This datasets have diverse characteristics and purposes, according to the research or need that generated them. Some datasets are taken from realistic conditions, such as YouTube videos, movie fragments, Olympic sports, surveillance video recordings and so on. Other datasets, mainly the early ones, belong to very controlled ambient conditions in which actions are performed by actors and the number of actions is reduced.

It is important to have easily accessible datasets, as this allows to make comparative studies between the proposed algorithms without bias to a particular data collection. This ability to be compared, gives higher confidence to the data and conclusions obtained respecting the performance of each algorithm.

A brief description of the most representative datasets for human action recognition is given next. For a detailed description the reviews presented in [8, 18] cover the datasets generated until the year 2013. Here are also presented the most recent datasets which aim to increase the scale respecting number of samples and action classes.

Among the first action recognition datasets are Weizmann and KTH. Weizmann is the first dataset for human action recognition. It is made of 10 actions performed by 9 actors in a controlled environment with static background. Along with the Weizmann dataset, KTH is one of the simplest and most used for the action recognition task. It is made of only 6 actions performed by 25 actors. Opposed to the Weizmann dataset, complexity is increased by making shots in 4 different contexts: Outdoors, outdoors with camera movement, outdoors with varying clothing and indoors. These datasets have been widely used and the performance reported on them is close to perfect recognition.

One of the limitations of the first datasets was their lack of realism as the environments were controlled and actions were performed by actors. To address this limitation, new datasets have been generated from realistic videos mostly from web video sharing platforms. Some of these first initiatives include UCF Sports and Hollywood datasets. The actions of the UCF Sports dataset belong to sports and where recorded from ESPN and BBC broadcasts so they have higher complexity and realism with respect to the KTH and Weizmann datasets. It has 9 action classes with a variable number of samples per action. The Hollywood dataset presents a high degree of realism and complexity as it was generated from movie sequences. It is made of 8 action classes taken from 32 movies.

The previously mentioned datasets improved the realism of the video sequences that composed them. But they still had a limited number of samples and action classes. This limitation has been given attention by some of the most recent datasets whose aim has been increasing scale.

Some of the first steps towards increasing scale include the UCF50 [49] and HMDB51 [36] datasets which have almost the same number of action classes. UCF50 is the evolution of a moderate size dataset. The original dataset (UCF11) has 11 action classes taken from YouTube videos and present uncontrolled environments (camera motion, varying execution speeds and viewpoints, etc.). This new dataset extends the number of actions to 50 and each action has at least 100 samples. The samples are subdivided in 25 groups. A group corresponds to segments taken from the same source video. Even though the segments come from the same source video, the group corresponds to scale or viewpoint variations. The ground truth for this dataset are the action and group labels. The HMDB51 dataset has 51 action classes with at least 101 sample segments. The videos where taken from movies and public websites such as YouTube and GoogleVideo. Actions are grouped in 5 categories: General facial actions, facial actions with object manipulation, general body movements, body movements with object interaction and body movements with human interaction.

Until this point, human action datasets are still far from reaching the same scale as image datasets which have already crossed the million of samples barrier. But the most recent datasets are getting close to reach that scale also. A first attempt has been made by expanding the UCF50 dataset to obtain the UCF101 [60] dataset with 101 action classes and a total of 13,320 sample videos with temporal segmentation. Based on UCF101 and using the same number of action classes, the THUMOS [24, 25] challenge has provided new video sequences with the difference of being temporally untrimmed. This implies that a sample video sequence may have more that one action occurrence and also actions not belonging to the 101 action classes. Nevertheless, the number of video samples provided are still below

40,000.
The Sports1M [33] dataset crossed the 1 million samples barrier and also increased the number of categories to more than 400, with the restriction of action classes belonging only to sports. The ActivityNet [14] dataset is an initiative to achieve scale close to the image datasets with the help of crowdsourcing. One interesting property of the dataset is that its collection process allows the dataset to keep growing in number of samples and action classes and human annotators increase the confidence of the labels assigned, opposed to Sport1M which has weak labels generated by an automatic process. The current version of the dataset has 203 activity classes evenly distributed among different categories based on an ontology. This is another advantage of the dataset which has been generated with the aim of avoiding bias towards a particular activity category. This type of datasets are the most appropriate for a retrieval system due to the diversity of categories, which allows the system to solve wider information needs of the users.

2.2 Features

Video features have been influenced by the work on image features. First approaches attempted to apply image features by adding the temporal dimension to the features, as motion had shown to be an important factor in describing actions. Due to the impact that CNN [35] features have had in image analysis tasks on the last years, video features proposals involving them have also been presented making evident the influence in image analysis advances on video analysis.
Interest point detectors were proposed by [39, 11] to select the most interesting video regions to extract features. The detectors focused on regions of high variability of motion and intensity. In [38] Histograms of Gradient and Flow (HoG/HoF) were used to encode static and motion information of movie videos.
As many proposals of features were presented to recognize human actions, an evaluation was carried in [68] to determine which were the most important aspects affecting performance. Along with interest point detectors, dense sampling of features was evaluated and it achieved better performance than sparse sampling with detectors. This motivated the author of the evaluation to propose new features based on dense sampling.
The proposed features are Dense Trajectories (DTF) [66] and Improved Dense Trajectories (IDT) [67]. Both are based on dense tracking of pixels and extracting features along the tracked trajectories. Tracking is based on dense optical flow with median filtering to estimate the position of pixels on consecutive frames. Trajectories are limited to a fixed length and filtered according to autocorrelation [53]. A volume is selected around trajectories and subdivided in the spatial and temporal dimensions. Four features are extracted and accumulated in each of the subdivisions. The four descriptors are Histogram of Gradients (HoG), Histogram of Flow (HoF), Trajectories and Motion Boundary Histogram (MBH). The descriptors encode appearance and motion information, among them MBH is the one

that performs the best when no fusion is performed. IDT improves over DTF by filtering flow based on an homography estimation and optional filtering by human detection. These features achieve state of the art performance on datasets like UCF101, HMDB51 and THUMOS.

With the advances achieved by CNN features in object detection in images, some CNN feature proposals have been made for video. Different architectures were presented in [33] and evaluated in the UCF101 dataset. The performance achieved was above the baseline presented with the dataset but below the best results in [24] using IDT. The authors of [56] attribute this low performance to a setting where flow was not included in the network input forcing it to also learn motion. They propose to solve this problem by training a two stream network with an appearance stream and a motion stream were stacked flow is the input to the network. This proposal achieves competitive results to IDT in the UCF101 dataset.

Not only the features have played an important role on action recognition advances but also the encoding method applied. Again, ideas previously applied to images are also applied to videos. Instead of using the Bag of Features (BoF) encoding where a feature vocabulary is generated (K-means is a common choice for this purpose) and features are aggregated on a histogram of the vocabulary according to the closest word, high dimensional encodings are used paired with linear models. One of this encodings is Fisher Vectors [23] whose main idea is to combine generative and discriminative models and exploit the benefits of each approach. The first step is to obtain a generative model of the features and Gaussian Mixture Models (GMM) are a common choice for this. Then, the gradients of a variable number of features from a sample entity (image, video, etc.) with respect to the generative model parameters (mean and covariance for the GMM case) are accumulated. The accumulated gradients are stacked to obtain a fixed length representation of the entity. For the case of the GMM the length of the representation is given by $2KD$ where D is the input feature dimensionality and K are the number of components of the GMM.

2.3 Video retrieval

In [70] a video content classification is proposed between scripted and unscripted material. Video with script is made of recordings where a sequence of actions is established by a script like in news, movies, etc. While videos without a script follow less well defined structural patterns, such as sports or surveillance videos. The authors state that according to the type of video, different collection representation methods should be used. For the case of scripted content, the following structure may be used.

- Scene
- Group
- Shot

- Keyframe

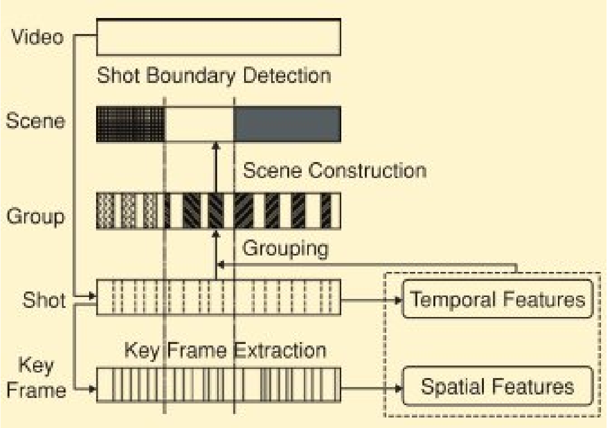


Figure 2.1: Representation of a scripted video. Taken from [70]

And in unscripted content, a representation is obtained focused on key aspects of visual and audio type, assuming these are the most relevant.

- Highlight group
- Highlight candidate
- Audio-Visual Markers
- Play/Break

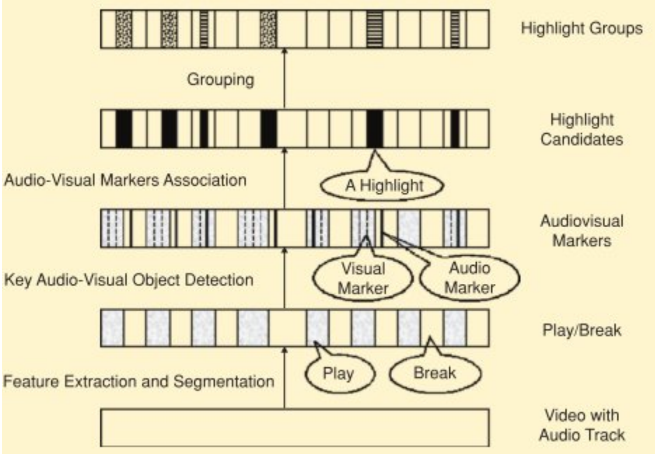


Figure 2.2: Representation of an unscripted video. Taken from [70]

A work of high importance and widely cited on the area of video retrieval is presented in [19]. In this work a theoretical analysis is made about the number and precision necessary in concept detectors in video to achieve automatic annotation and obtain a retrieval system with 65% MAP, equivalent to the MAP of the best text based web search engine at the time of the study. The work is based on an evaluation of the performance of a retrieval system using different numbers of concept detectors and the result is extrapolated to the expected MAP. The detectors used assume only 10% precision on concept detection and with this assumption the result is that 4000 concept detectors would be necessary to achieve the goal. It is also found that annotating only 10% of the collection would allow to find 90% of the most frequent concepts.

With the aim of dealing with the problem of the semantic gap, in [62] it is proposed to use external resources to obtain sample images for video search. The idea is to provide the user with a keyword based retrieval system and to use sources like DBPedia, Flickr and Google to perform the translation from keywords to images, which are then used by the retrieval system for comparison with video frames and determine the relevant ones to the keyword query.

In [57], a proposal is made to apply text retrieval techniques to video. This is made by first applying the Bag of Features technique to obtain an analogous representation to words in text (Bag of Words). To this representation, stopword removal is applied and then a TF-IDF weighting scheme is carried and experiments are performed with different dictionary sizes and similarity measures.

By means of genetic programming, local and global descriptors are fused in [1]. For the retrieval of videos focused on Computer Assisted Retinal Surgery, [37] uses color, texture and movement descriptors and Approximate Nearest Neighbors (ANN) for the retrieval under the Euclidean distance. The problem of retrieval with missing frames is tackled in [51] by means of PCA for dimensionality reduction and then sparse coding is used for the reconstruction. [69] proposes Words of Interest which is the use of the Bag of Words technique adding a filtering of words according to their spatial relations.

Retrieval focused on traffic surveillance videos is presented in [21] allowing querying by keywords, multiple objects and sketches. For this, vehicle trajectories are extracted and tags are generated with manual annotation over part of the collection to allow searching. The trajectories are grouped by means of spectral and hierarchical clustering and one activity is assigned to each cluster.

In [2] use of context along with video content is proposed for retrieval. The concept of context in this work makes reference to groups of similar videos and it is aimed at measuring grouping among the collection videos. The context is represented by measuring the similarity of each video with the rest of the collection. The similarity is calculated using the video content descriptor. The content descriptor is a probability distribution obtained from the intensity gradient over a temporal pyramid of the video sequence. As a similarity measure for the content descriptors the χ^2 divergence is used and the cosine similarity for the context descriptor. For the combination of individual similarities a sum and weighted product is proposed allowing to vary the influence of each descriptor on the retrieval.

2.4 Human action video retrieval

The work presented in [44] employs as a representation the output of 3D Gabor filters with max-pooling to simulate the human visual system. A GMM is trained with this representation and by means of sliding windows an approximate temporal location of the action is obtained.

In [29] Relevance Feedback is evaluated in human action retrieval. Relevance Feedback consists of generating an initial result over which an user gives feedback iteratively to the system about the relevant and irrelevant results so that the system may improve retrieval and generate new results with higher precision. The authors start from direct comparison of visual features to generate the initial ranking. User feedback is used to train a SVM which is able to discriminate relevant elements of the collection. The proposal uses a ABRS-SVM which helps to compensate for class imbalance and few training examples. The authors improve the experimentation in [28] using hard and soft clustering, different descriptors and representations such as Bag of Words, Vocabulary-Guided Pyramid Match and Spatio-Temporal Pyramid Match.

[9] makes use of position, velocity and color as descriptors. Clustering is performed over the representation using Hierarchical Mean Shift Analysis and comparison using the centroids.

Human behavior analysis inside a sales point is presented in [54]. The authors selected 6 states in which a person may be inside a market (Enter, Exit, Interested, Stand By, Inactive, Interaction) and the behavior is modeled by means of a Finite State Machine. For feature extraction 2 preprocessing steps were performed. First, background detection based on movement and classification with a Gaussian Mixture Model. Then, object tracking was carried, which was interpreted as humans interacting with the environment. Objects were represented by using Motion History Image, Accumulated Motion Image, local motion context based on optical flow and interaction context based on the person bounding box. these descriptors were used with a SVM to detect when a person interacted with the market racks.

The proposal made in [22] uses Hidden Markov Models for the retrieval of human actions. In that proposal, an additional data source is used for training the model, consisting of Motion Capture data. This data corresponds to information read from sensors, such as accelerometers, located on body parts of an actor. A special suit may also be used with markers to be tracked. The collected information is position, velocity and acceleration of the body parts. The body is divided in 4 sections: left arm, right arm, left leg and right leg. The objective is to carry a discriminative analysis of simultaneous actions, for example walking and waving. The queries are made by means of regular expressions instead of visual examples, the regular expressions correspond to the action performed by each body part.

3 An Evaluation of NMF Algorithm on Human Action Video Retrieval

Human action video retrieval is a useful tool for video surveillance and sports video analysis, among other applications. Previous work on image retrieval tasks has shown that latent semantic methods are an effective way to build a high-level representation of data to discover implicit relations between visual patterns, achieving a significant improvement on these tasks. The current paper evaluates the applicability of Non-Negative Matrix Factorization (NMF), a latent semantic method, on human action video retrieval. Experiments are carried out on common human action recognition datasets using state-of-the-art descriptors. We focus on evaluating the query by example approach i.e. only videos are used as queries. The performance of the method is compared against classic direct matching between video features. ***This chapter corresponds to an article submitted and presented at the "Simposio de Tratamiento de Señales, Imágenes y Vision Artificial 2013" organized by the Universidad Antonio Nariño, and is part of the event proceedings indexed by IEEE. It is joint work with MSc Jorge Vanegas and PhD Fabio González.***

3.1 Introduction

Video is an important mean to record information. It has the ability to combine different information modalities like visual, audio and text content. This inherent multi-modality brings the need of developing video analysis algorithms that exploit this content richness.

Recent reports[12] show that video is an important part of mobile data traffic (36%) and is fore-casted to increase its share to 41% for 2018. This important role of video content can be attributed to the growth of digital camera availability on mobile devices. The trend is also fueled by the appearance of media sharing sites and social networks which have received wide acceptance by the public. The current availability of video content and its predicted growth give place to customer needs like browsing and searching on huge video databases among others like categorization and summarization. This needs are addressed by the field of video analysis. Within this field, video retrieval is an useful tool to achieve the browsing and search needs.

As most of the video content is generated by humans, we assume that most of it also features humans performing a wide range of actions and activities. This assumption allows us to believe that human actions may be an important discriminative factor to perform video retrieval. Besides, a wide range of applications could benefit from these types of systems. Some of these are surveillance, entertainment, sports[44] and marketing[55]. Nevertheless, this is not an straightforward task. Like many video analysis tasks, factors like occlusion, viewpoint, speed and scale variance, noise, etc. make it a hard task.

In this paper the performance of NMF-based algorithms on the task of human action video retrieval is evaluated. The NMF method purpose is to find a space in which patterns are easily spotted. This is achieved through a matrix factorization of the input data that performs a dimensionality reduction, generating latent factors (or clusters), as well as a matrix indicating the degree of belonging of each sample to each latent factor. The paper is organized as follows: a brief summary of previous work is given in section 3.2; then, in section 3.3, the proposed method is described; next, the experimental setup is explained in section 3.4 followed by the corresponding results in section 3.5; finally, concluding remarks are given in section 5.5.

3.2 Previous Work

On the field of human action recognition, [43] proposed using unsupervised learning by means of pLSA. In this approach, the number of latent topics of the pLSA method where chosen according to the number of actions present on the datasets used for experimentation. The unsupervised character of the approach makes it useful to tackle harder problems with little or no annotation information. This method keeps a close resemblance to NMF [10] and is thus part of the motivation to our proposal.

The work presented in [26, 29, 28] focuses on relevance feedback as the way to achieve state-of-the-art performance on the task of human action retrieval on video for different action recognition datasets. The feedback is performed on the retrieval results from direct matching between the visual representation of each video using the BoF model. User feedback is used to train a SVM to provide new rankings for the user query.

The notion of context along with content to describe a video is introduced in [2]. Here, context is understood as a measure of similarity among the videos of the collection which generates a new descriptor for the video with dimensions equal to the size of the collection. Context in this way can be understood as a mean to group or cluster similar videos. As two descriptions of a video are obtained (content and context), their measured similarities must be fused to perform final similarity and retrieval. This is done using weighted sum and product of each similarity.

The NMF algorithm has been successfully used for image retrieval. In [6] the NMFA method was used for indexing on the Corel5K dataset and it achieved better results than NMFM, direct visual matching and SVD. NSE was applied to histology image indexing in [64]. Both works used a visual modality and a textual or semantic modality made of the labels or categories assigned to each image.

3.3 NMF Algorithm on Human Action Video Retrieval

The non-negative matrix factorization (NMF) algorithm aims to produce a factorization for a given matrix X under the constraint of non-negativity. The factorization can be interpreted as finding a space to represent the data given by X in which W denotes the latent factors and H the projection over those factors.

$$X \approx WH \tag{3.1}$$

$$W, H \geq 0 \tag{3.2}$$

3.3.1 NMF algorithms

In the case of multi-modal problems, some variations have been proposed. The modalities presented here are made of a visual modality X_V and a text modality X_T. The evaluated algorithms look for relations between this modalities by projecting data to a common latent space in which patterns are easily noticed and compared.

The dimensions of the X_V, X_T matrices are given by the size of the modality descriptor and the number of samples. Assuming n as the visual representation dimension, m as the textual representation dimension and l as the number of samples, we have $X_V \in \mathbb{R}^{n\times l}$, $X_T \in \mathbb{R}^{m\times l}$.

NMF Asymmetric

The NMF asymmetric (NMFA) [6] algorithm builds a latent representation by the decomposition of the text matrix

$$X_T = W_T H \tag{3.3}$$

which is used to find a latent space basis for the visual data

$$X_V = W_V H \tag{3.4}$$

The visual latent space basis W_V is used to project queries with only visual data to the latent space for comparison with the database latent representation H.

Non-negative semantic embedding (NSE)

This variation of the NMF algorithm aims to embed the visual data on a space generated by the text data

$$X_V = W X_T \tag{3.5}$$

Here, the problem is to find the latent space basis, W, which is used to embed queries, with only visual data, in the semantic space. The semantic representation of queries is used to retrieve semantically similar videos from the database.

NMF for Content-based Video Retrieval

In order to apply NMF to human action videos, the visual content of the video, X_v, must be represented using features that characterize spatial-temporal information. These features are described in the following subsection. The textual content, X_t, corresponds to a matrix with the labels of the human actions contained in each video.

Video representation

Following the work in [38, 39], a space-time interest point (STIP) detector was used to select the points with high spatial and time variation in order to perform feature extraction. The extracted features are HOG/HOF histograms. These features and STIP were extracted using the code provided in [1]. After feature extraction a BoF model was used to represent each video.

Video Indexing and Search

Having obtained the latent factors represented by $W \in \mathbb{R}^{n \times r}$, new samples can be projected to the generated semantic space. This is done to perform comparison in the latent space where patterns have been found. For a given query, x_v, with only visual data available, a semantic representation can be obtained solving the following system

$$x_v = Wh \tag{3.6}$$

This column vector $h \in \mathbb{R}^{r \times 1}$ is compared, under a selected similarity measure, against all the elements of the index, which is $H \in \mathbb{R}^{r \times l}$ in the NMFA case and X_T in the NSE case. r corresponds to the number of latent factors and is a parameter for the NMFA algorithm but fixed and equal to m for NSE. After applying the similarity measure, a vector $s \in \mathbb{R}^l$ holds in the i-th position the similarity between the query and the i-th database sample. This vector indicates the relevance of each sample for the query.

3.4 Experiments

3.4.1 Dataset

The selected dataset is the UCF50 human action dataset[49]. This dataset is an extension to the UCF11 dataset[52] and contains 50 different actions with realistic settings taken from

[1] http://www.di.ens.fr/~laptev/download/stip-2.0-linux.zip

YouTube videos. Every action is guaranteed to have at least 100 sample video segments. The sample video segments are subdivided in groups and every group is made of segments from the same source video. The only information provided with the dataset is action label and group correspondence for each segment.

This dataset aims to provide a more realistic setting compared with previous datasets and a wider number of actions. The number of actions contained make it a challenging dataset both for recognition and retrieval tasks.

3.4.2 Experimental setup

As mentioned earlier, the videos are represented visually using the BoF model. First, k-means was used to generate feature clusters over a sample of 10% of the whole dataset representation. Then, for each video segment and using the generated codebook, a histogram was created counting the occurrence of each visual word on the segment. Different number of clusters were generated and tested, resulting in different video representation dimensions. The selected k values were 500, 1000, 2000 and 3000.

The dataset was divided according to the suggestions made in [49], which are using Leave-One-Group-Out (LOgO) cross-validation. This is made to avoid a performance increase by having segments from the same source video in training and testing partitions. As there are at least 25 groups per action, [49] suggest making 25 runs of LOgO. Using this setup, for every cross-validation run, 24 groups of the 50 actions are used as database, each group having in average 4 video segments, and for queries, 1 group of each action is used to retrieve similar videos from the database. An average of $50\,actions \times 1\,\frac{group}{action} \times 4\,\frac{segments}{group} = 100\,segments$ are used for queries in each LOgO run.

The dataset was divided in two modalities: a visual modality made of the video histogram and a textual or semantic modality made of a 50-dimensional vector with a value of 1 in the bin corresponding to the action present in the video and 0 on the others.

The index is built using both modalities to find the latent factors that exploit the richness of the semantic representation, but as the query by example strategy is used, query videos are represented only by the visual description generated by the BoF method and no textual information is available. This way, no previous labeling is needed on the query video.

For the selected algorithms (visual matching, NMFA and NSE), experiments were performed using the 4 values for cluster number (k-value). In all of the experiments the histogram intersection similarity measure was used. The selected performance measures are MAP and P@10 which are averaged over the 25 runs of LOgO. The number of latent factors for the NMFA algorithm were 50. This value was selected to match the number of actions.

3.5 Results

Table 3.1 shows the MAP and P@10 values for each value of cluster size or visual histogram dimension.

Table 3.1: Algorithm performance comparison

		k-value			
Algorithm	Measure	500	1000	2000	3000
Visual matching	MAP	0.106	0.112	0.118	0.122
	P@10	0.277	0.289	0.300	0.307
NMFA	MAP	0.346	0.353	0.396	0.411
	P@10	0.273	0.276	0.320	0.339
NSE	MAP	0.381	0.410	0.442	0.463
	P@10	0.315	0.345	0.376	0.397

This information is also shown on Figure 3.1 and 3.2. It can be seen that increasing the visual histogram dimension improves the values for both of the performance measures. The improvement can be attributed to a greater discriminating property of a bigger visual vocabulary size. Further increase of the vocabulary size could achieve better results with the drawback of longer computing times. The effect of this parameter is more evident on the values of the performance measures for NSE algorithm. For this algorithm there is always an steady increase, while for the other two algorithms the values remains almost the same with an increase of vocabulary size.

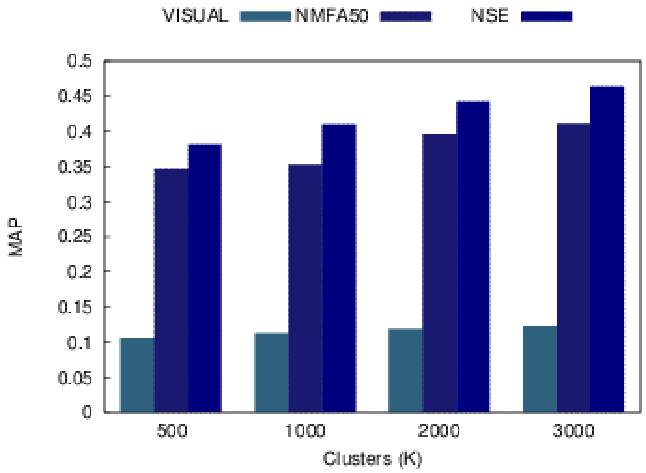


Figure 3.1: MAP comparison

The best performing algorithm based on MAP measure is NSE. It is closely followed by NMFA but always with lower values. The worst performing algorithm is direct visual matching. The gain achieved with the proposed algorithms over visual matching is around 4 times. This gain confirms the applicability of the method to perform human action retrieval.

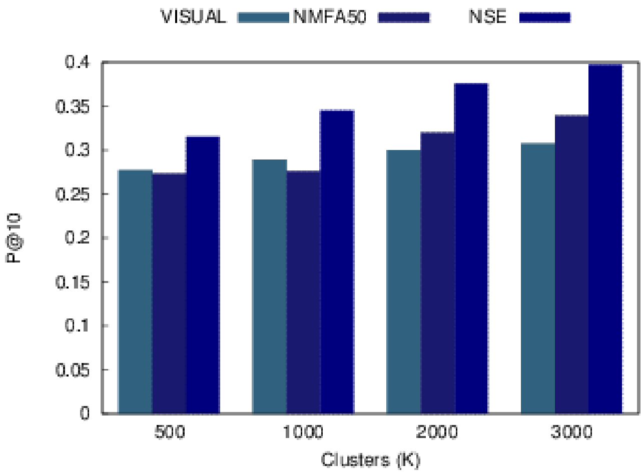


Figure 3.2: P@10 comparison

For the case of early precision, measured with P@10, there is not a strong gain as with MAP measure. Visual matching presents better results in early precision than those of MAP. NMFA achieves a better P@10 than visual matching but only for the 2000 and 3000 visual dictionary size. Its performance is slightly lower for the 500 and 1000 dictionary size. Finally, as in the case with MAP, the NSE algorithm achieves the best P@10 results for every vocabulary size.

3.6 Conclusions

Different algorithms based on NMF are proposed to perform human action retrieval on video. To compare the performance of the proposed method, visual matching was used as a baseline. The obtained results show that the proposed method achieves a significant improvement over the selected baseline. This improvement is stronger for the MAP measure used to evaluate the performance of the methods. The best performing method is NSE with almost 4-fold increase of MAP relative to visual matching.

3.6.1 Future work

Among the possible points to improve is the selected visual descriptor used to represent codewords, such as those in[34][49]. The algorithms could be also tested on the HMDB dataset[36] which resembles closely the UCF50 dataset on scale factor. Using different similarity measures may improve the obtained performance.

Acknowledgement

This work was partially funded by the project Anotación Automática y Recuperación por Contenido de Imágenes Radiológicas Usando Semántica Latente, number 110152128803 by Convocatoria Colciencias 521 de 2010.

4 Online Multimodal Matrix Factorization for Human Action Video Indexing

This paper addresses the problem of searching for videos containing instances of specific human actions. The proposed strategy builds a multimodal latent space representation where both visual content and annotations are simultaneously mapped. The hypothesis behind the method is that such a latent space yields better results when built from multiple data modalities. The semantic embedding is learned using matrix factorization through stochastic gradient descent, which makes it suitable to deal with large-scale collections. The method is evaluated on a large-scale human action video dataset with three modalities corresponding to action labels, action attributes and visual features. The evaluation is based on a query-by-example strategy, where a sample video is used as input to the system. A retrieved video is considered relevant if it contains an instance of the same human action present in the query. Experimental results show that the learned multimodal latent semantic representation produces improved performance when compared with an exclusively visual representation.

This chapter presents the work submitted and presented at the "Content Based Multimedia Indexing 2014" conference, and is part of it's proceedings indexed by IEEE. It is joint work with MSc Jorge Vanegas and PhD Fabio González.

4.1 Introduction

Thanks to the widespread use of smartphones with capabilities for recording high definition images, videos and audio, multimedia content has experienced an exponential growth. Social networks have also contributed to this growth as they have provided the means to publish and share such content. This trend exposes every cyber-citizen to a myriad of content everyday. Such overwhelming amount of information is part of the Big Data phenomenon, which brings forth the need for efficient ways to manage and analyze the ever-growing content corpora. Among those needs, is the one of information search. One of the main challenges in building an effective search and retrieval system is to deal with the semantic gap problem, i.e., the fact that retrieving visually similar content may be unsatisfactory since, in general, these results could be semantically unrelated to the query. A way to deal with this problem is to use multimodal data to enrich the data representation with relations not present in a single modality. Another challenge is to deal with the large, and continuously growing, multimedia collections. This means that any proposed method to deal with the aforementioned problem has to scale to large multimedia collections.

In this paper, we focus on the problem of identifying and searching videos containing particular human actions. This is an important problem since video is one of the fastest growing types of data in the web. Assuming that a good portion of the videos published on the web, many of them published on social networks, involve human actions[15], being able to analyze which actions/activities appear in a video would be an effective mean to organize video collections.

This paper presents a method for online, multimodal video indexing. The method uses a multimodal latent semantic embedding based on matrix factorization. The factorization is performed on an early fusion of the modalities, aimed at finding a common latent representation for a collection item. This representation is used to built the collection index, which is compared with the latent representation of a query. Experimental evaluation is carried on a query by example setup, where queries provide visual data only. A dataset of human action videos is used as collection for experimentation. The two main contributions of this work are the extension of the method to more than two modalities and experimental evaluation on a video collection.

The rest of the paper is organized as follows: in Section 4.2 a brief presentation of previous work is given; in Section 4.3 the proposed method is described; Section 4.4 presents the experimental evaluation and results; finally, Section 4.5 discusses the conclusions and future work.

4.2 Previous work

Human action retrieval has been addressed in [27, 30, 28] focusing on relevance feedback as a way to improve retrieved results by visual matching only. Evaluation of the proposed method was carried on realistic YouTube videos and the Hollywood dataset, using as modalities the visual content and the action label of each video.

Non-negative matrix factorization (NMF) is a dimensionality reduction technique which factorizes a given non-negative matrix $X \in \mathbb{R}^{m\times n}$, in two non-negative low rank matrices ($W \in \mathbb{R}^{m\times r}$, $H \in \mathbb{R}^{r\times n}$) satisfying the relation $(n+m)\,r < nm$. NMF with multiplicative update rules is presented in [41]. Using multiplicative rules rather than additive ones guarantees satisfying the non-negativity constraint. This constraint is shown in [40] to be useful to obtain a parts-based representation of data in the context of image reconstruction and textual semantic topic modeling.

The work presented in [5] presents an exhaustive evaluation of NMF in the context of multimodal indexation and annotation of image collections presenting the mixed and asymmetric variants of the algorithm. A comparison against singular value decomposition (SVD) is performed, and the proposed methods show the effectiveness of building a latent representation from multimodal information, made of the visual content and associated tags or labels. PLSA is employed in [43] to build an unsupervised classifier of human action videos. On [13], NMF was used to solve the problem of action recognition on still images with different views, which resemble multimodal information. This methods are based on a latent space representation of data, which in the context of matrix factorization is given by the H matrix.

Stochastic gradient descent (SGD) is described in [4]. This algorithm allows to reduce training time given that learning is carried on a large scale dataset. That work shows that using SGD, a single pass over the training data is enough to reach an asymptotically better empirical risk than non-stochastic gradient descent. An online matrix factorization method is proposed in [7] which solves the factorization using additive rules and SGD. The method was evaluated on a retrieval setup on image collections and compared against batch NMF methods. Based on that work, [45] address the multilabel annotation task of images. The latter and former methods focus on modeling the latent space using two data modalities.

Early and late fusion is evaluated for semantic concept detection in video collections in [59]. According to that work, late fusion generated best results in most of the concepts but with a small difference to early fusion.

The current work proposes an extension of online matrix factorization to more than two modalities by means of early fusion of the multiple modalities information. It also evaluates the algorithm on a different domain, which is human action videos.

4.3 Multimodal latent semantic embedding for video indexing

The idea behind latent semantic embedding is to find a space in which semantic structure of data is modeled. When multiple modalities are involved, the space generates relations among the modalities. In the case of matrix factorization, a linear transformation relates the modality data to the latent space as illustrated in Figure 4.1.

Let's assume that each video in the collection is represented by N different modalities and each modality i is represented by m_i features. Then, the video collection may be represented by a matrix $X \in \mathbb{R}^{m \times l}$:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \tag{4.1}$$

where, l is the number of videos in the collection, $m = \sum_{i=1}^{N} m_i$, and $X_i \in \mathbb{R}^{m_i \times l}$ is the collection representation of the i-th modality.

We assume that all the modalities describing videos may be represented (embedded) in a

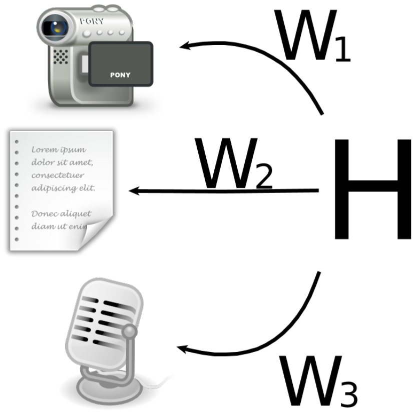


Figure 4.1: Latent semantic embedding based on matrix factorization aims to obtain a latent representation H related to each modality by means of a linear transformation W. The diagram models the projection from the latent space to each modality e.g. video, text, audio.

common latent semantic space. The goal of the method is to find this common representation. This is accomplished by finding different linear transformations W_i, which generate each modality collection representation from the common latent representation, $H \in \mathbb{R}^{r \times l}$, where r is the dimension of the latent semantic space. So, we expect that:

$$X = WH \tag{4.2}$$

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_N \end{bmatrix} H = \begin{bmatrix} W_1 H \\ W_2 H \\ \vdots \\ W_N H \end{bmatrix} \tag{4.3}$$

Our problem is now, how to find the W_i transformation for each modality and corresponding H . This is accomplished by solving the following optimization problem:

$$\begin{aligned} \min_{W,H} \mathcal{L} = \min_{W,H} \sum_{i}^{N} \left(\alpha_i \parallel X_i - W_i H \parallel_F^2 + \lambda \parallel W_i \parallel_F^2 \right) \\ + \lambda \parallel H \parallel_F^2 \end{aligned} \tag{4.4}$$

where α_i controls the importance given to the $i-th$ modality in the optimization, λ controls the amount of regularization of the parameters, aimed at avoiding overfitting.

To solve the minimization problem, the objective is differentiated with respect to W_i and H

and equated to 0.

$$\begin{aligned}\frac{\partial\left(\mathcal{L}\right)}{\partial W_i} &= 0 \\ &= \sum_{i}^{N}\left(\lambda W_i - \alpha_i\left(X_i - W_i H\right)H^T\right)\end{aligned} \tag{4.5}$$

$$\begin{aligned}\frac{\partial\left(\mathcal{L}\right)}{\partial H} &= 0 \\ &= -\sum_{i}^{N}\left(\alpha_i W_i^T\left(X_i - W_i H\right)\right) + \lambda H\end{aligned} \tag{4.6}$$

A closed solution for H can be obtained by equating the corresponding gradient to zero 4.6, and solving for H

$$\begin{aligned}H_\tau = &\left(\lambda I - \sum_{i}^{N}\left(\alpha_i W_i^T W_i\right)\right)^{-1} \\ &\cdot\left(\sum_{i}^{N}\left(\alpha_i W_i^T\left(X_i\right)\right)\right)\end{aligned} \tag{4.7}$$

$$H_\tau = \left(\lambda I - W^T diag\left(\alpha\right)W\right)^{-1} W^T diag\left(\alpha\right)X \tag{4.8}$$

where $diag\left(\alpha\right)$ is a diagonal matrix of the modality weights, each repeated a number of times equal to the feature dimension of the $i-th$ modality. The τ subindex indicates the current iteration involving the update of the parameters.

And the gradient descent update formula for W is

$$\begin{aligned}W_i\mid_{\tau+1} &= W_i\mid_\tau - \gamma\frac{\partial L}{\partial W_i} \\ &= W_i\mid_\tau + \gamma\left(\alpha_i\left(X_i - W_i\mid_\tau H_\tau\right){H^T}_\tau - \lambda W_i\mid_\tau\right)\end{aligned} \tag{4.9}$$

$$\begin{aligned}W_{\tau+1} = &\left(1-\gamma\lambda\right)W_\tau + \gamma diag\left(\alpha\right)XH_\tau^T \\ &-\gamma diag\left(\alpha\right)W_\tau H_\tau H_\tau^T\end{aligned} \tag{4.10}$$

Until now, we have stated the update equations for a batch method. To introduce SGD we are going to process a single sample at a time, which implies changing H by h and X by x which are the latent representation and concatenation of modalities for a single training input to the system, respectively.

Following the work presented in [7], the SGD algorithm may be extended in order to process not a single sample but a small batch (minibatch). This minibatch size becomes another parameter of the algorithm, and setting it equal to the size of the training set corresponds to the non-stochastic algorithm.

According to [4], algorithm convergence is guaranteed if the learning rate γ is updated by the following formula

$$\gamma_\tau = \frac{\gamma_0}{1 + \lambda\gamma_0\tau} \tag{4.11}$$

To project queries containing information from the $i-th$ modality only, equation 4.12 is used, which is obtained by solving $\min_{h_q} \parallel x_{qi} - W_i h_q \parallel_F^2 + \lambda \parallel h_q \parallel_F^2$

$$h_q = \left(\lambda I + W_i^T W_i\right)^{-1} \left(W_i^T x_{qi}\right) \tag{4.12}$$

The algorithm has 6 parameters to tune:

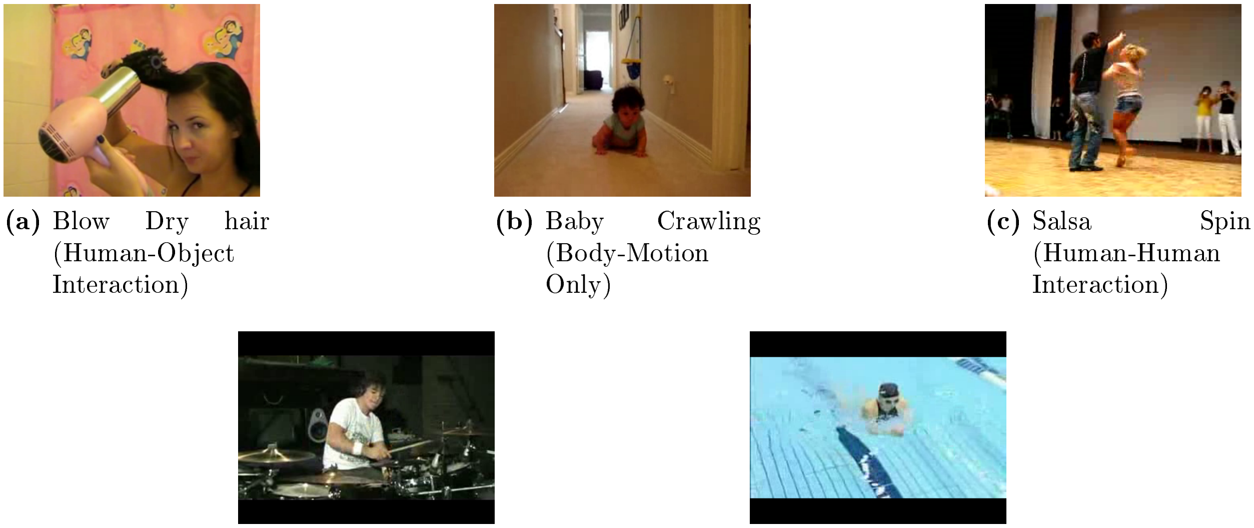
- λ, the regularization
- γ_0, the initial value of the learning rate
- α, the weights for each modality
- the number of epochs, or iterations over the full training data
- the minibatch size, number of samples used at each update
- r, the number of latent factors

4.4 Experiments

The proposed multimodal semantic embedding method is evaluated on a video retrieval task focused on human actions as the information need. The collection index consists of the latent space representation of the database H, which is built from different modality groups to assess the contribution of each modality to the index quality. Queries provide unimodal visual information which is projected to the latent space. Once the query is represented on the latent space, comparison against the index is performed under a similarity measure to obtain ranked results. Relevance judgement is based on a collection item sharing the same action label as the query.

4.4.1 UCF101 dataset

This dataset [60] focuses on human action recognition and provides the current largest number of actions for that task [18]. The clips from this dataset come from 101 actions sampled from YouTube videos and amount to a total of 13320 clips with more than 100 samples per class. The actions are grouped in the following 5 types: Human-Object Interaction, Body-Motion Only, Human- Human Interaction, Playing Musical Instruments, Sports.
The visual representation of the collection is based on a bag of words (BOW) of dense trajectory features (DTF) [66] features released as part of the “THUMOS Challenge: Action Recognition with a Large Number of Classes” [24]. The DTF features comprise histogram



(a) Blow Dry hair (Human-Object Interaction)

(b) Baby Crawling (Body-Motion Only)

(c) Salsa Spin (Human-Human Interaction)

(d) Drumming (Playing Musical Instruments)

(e) Breaststroke (Sports)

Figure 4.2: Sample frames for actions of each of the 5 types

of gradients (HOG), histogram of flow (HOF), motion boundary histogram (MBH) and trajectory descriptors extracted along densely sampled trajectories. These trajectories are limited to an specific number of frames to avoid trajectory drift. Densely sampling instead of interest point-based feature extraction was shown to yield an improvement in human action recognition in [68]. The codebook size for each of the descriptors is 4000. Only the MBH descriptor was used as within those features, it yields the best results for action recognition by itself [66], this was made to reduce the time complexity of the experimentation due to the high number of dimensions of the descriptors.

Along with the aforementioned features, the dataset has the corresponding action label for each clip and a mapping from actions to attributes [24], which provide a granular representation of the action. There are 115 different action attributes, which were manually assigned to each action class. This provides a representation of each class in a 115-dimensional binary vector. The mean attribute occurrence per action is 32. Attributes are divided in 14 groups: Body Part Articulation-Hand, Object, Body Part Articulation-Feet, Body Part Articulation-Arm, Number of People, Body Parts Visible, Body Motion, Body Part Articulation-Head, Outdoor, Body Parts Used, Body Part Articulation-Torso, Body Part Articulation-Leg, Indoor, Posture. The Outdoor and Indoor groups are also attributes by themselves. Table 4.1 shows some actions with a sample of its corresponding attributes.

The dataset was partitioned in training, validation and test subsets. The partitions used correspond to those defined by [24] for the classification task. For that task, three different splits were defined specifying a test set and a training set and each test set is mutually exclusive with the others. Taking advantage of this property, test partitions of splits 1 and

Table 4.1: Five actions with a sample of its corresponding attributes

Action	Attributes
Blow Dry Hair	Body Part Articulation-Arm = One_Arm_Stretched, Body Parts Visible = One_Hand, Body Part Articulation-Head = Tilted_Position, Number of People = One
Drumming	Body Motion = Flipping, Posture = Standing, Number of People = One, Body Part Articulation-Arm = Two_Arms_Motion, Body Part Articulation-Arm = Synchronized_Arm_Motion, Body Part Articulation-Leg = Two_Legs_Stretched
Salsa Spins	Body Part Articulation-Leg = Fold_Unfold_Motion, Body Parts Used = Foot, Body Part Articulation-Head = Straight_Position, Body Parts Used = Arms, Body Parts Used = Hands, Posture = Sitting, Body Part Articulation-Arm = Two_Arms_Bent, Body Parts Visible = Upper Body
Baby Crawling	Body Part Articulation-Head = Facing_Down, Outdoor = Grass, Outdoor = Sky, Indoor = Home, Body Parts Visible = Two_Hands, Number of People = One, Indoor
Breaststroke	Indoor = Home, Body Parts Visible = Upper Body, Body Part Articulation-Feet = Touching_Ground, Body Part Articulation-Head = Facing_Front, Body Parts Visible = Head_Closeup

Table 4.2: Size statistics of the partitions used for training, validation and testing

Partition	Size	Percent
Test	3783	28.4%
Validation	3734	28%
Training	5803	43.6%
Total	13320	100%

2 were used as test and validation queries, respectively. The remaining samples were used as training set. A detailed description of the size of each partition is given in Table 4.2.

4.4.2 Setup

The performance is measured in terms of mean average precision (MAP) and precision for the first 10 results (P@10). Algorithm parameters were tuned using cross validation over the training and validation partitions using a grid-search strategy. The best performing parameters, according to MAP, were then used to train on the union of training and validation partitions and reported results were calculated on the test partition. All the data modalities were normalized using L_2 norm. The partitions respect the guideline for the dataset of leaving samples from the same group in the same partition, as a group shares context and leaving instances of the same group in different partitions may yield an artificial performance gain.

Training is performed using combinations of the modalities available i.e. visual, attributes and actions. As the baseline is visual retrieval, multimodal combinations always include the visual modality. This generates 4 different modality groups for training:

- visual
- visual+action
- visual+attributes
- visual+attributes+actions

For *visual, visual + actions* and *visual + attributes*, the same set of parameters were explored. For the case of *visual + attributes + action*, there was a change only on the weight used for each modality. The λ, γ and latent factors explored are the same used in all modality groups. For just visual embedding, a weight of 1 is used as there is just one modality.

The selected parameters for each modality group used for testing are presented in Table 4.3 The table shows a pattern relative to modality weights. Visual weight is small relative to the weight given to the other modalities. This is due to the action-based relevance judgement, as the attributes and action modalities are directly related to the relevance judgement. Another parameter of the algorithm is the number of epochs, or iterations over the complete training

Table 4.3: Selected parameters for each modality group in training

Group	Latent factors	λ	γ_0	Visual weight	Actions weight	Attributes weight
Visual	1500	0.001	0.0001	1.0	N/A	N/A
Visual + Action	750	0.0001	0.0001	0.2	0.8	N/A
Visual + Attributes	500	0.001	0.001	0.2	N/A	0.8
Visual + Attributes + Action	500	0.001	0.001	0.2	0.2	0.6

Table 4.4: Performance measures for each modality group

Group	MAP	P@10	MAP Improvement	P@10 Improvement
Visual	0.1291	0.3181	N/A	N/A
OMF Visual	0.1457	0.3415	12.86%	7.36%
OMF Visual + Attributes	0.4740	0.4768	267%	49.9%
OMF Visual + Action	0.4771	0.4474	270%	40.6%
OMF Visual + Attributes + Action	0.5089	0.5060	294%	59.1%
QBSE	0.6931	0.6400	436%	101%

data, and minibatch size, number of samples used for every model update. The epochs parameter was fixed at a value of 4 for training and testing, and minibatch size was also fixed at 256 samples.

4.4.3 Results

The first baseline method is visual retrieval without any kind of semantic embedding. In this case, the visual representation of each document and query is compared using histogram intersection similarity. This similarity has shown good performance in retrieval when a BoW representation is used [50]. The second baseline is query by semantic example (QBSE)[58, 48] using a SVM with a histogram intersection kernel trained with the MBH descriptor and dot product as similarity measure in the semantic space generated by the SVM. For the case of comparing representations obtained using the multimodal semantic embedding, a dot product similarity is used as the latent representation is not constrained to have positive values, making histogram intersection inappropriate for that representation.

Performance results are shown in Table 4.4. Along with the performance measure, the relative improvement over the baseline is also presented in the table. It can be seen that using attributes gives an improvement in precision over using action labels. Besides, using the three modalities gives an improvement in both performance measures. Even though a performance gain is obtained with attributes instead of actions, it is relatively small. This can be attributed to the fact that a particular combination of attributes determine an action. Nevertheless, for the information retrieval task, attributes provide a fine-grained description

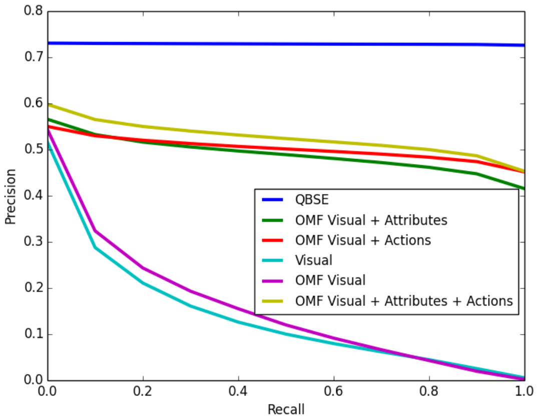


Figure 4.3: Interpolated precision recall for the different modality groups

of the actions, which allows to discover relations among collection items, not visible only with the action information.

Figure 4.3 shows the interpolated precision-recall curve for the different setups. It can be seen that visual features based retrieval has a high precision on early recall values, which is an expected result as some items of the same action may have visual similarities. But as recall increases, visual retrieval precision drops. Retrieval based on latent embedding of just visual features shows an improvement over the retrieval using raw visual features, but their behaviors are very similar. On the other hand, the curves for retrieval based on multimodal representations have higher precision values which drop slightly as recall increases. Among these, retrieval using the three modalities dominates the other curves.

The QBSE baseline outperforms the proposed method, this can be attributed to its supervised and discriminative nature. The proposed matrix factorization is an unsupervised method, yet it manages to generate a multimodal index that yields good results when using action performed on video as relevance judgement without being trained to optimize that goal. The QBSE method is trained aimed at discriminating this action and that translates in a superior performance when the action is the relevance judgement. A comparison of both methods using attributes as relevance judgement may provide interesting results when the concept detector keeps being trained with actions as the target concepts. A drawback of the type of concept detector used (HI-SVM) is the time and memory complexity of the kernel generation for a large number of training samples, even though this only needs to be computed once. Additionally, as the representation obtained with the proposed method outperforms the visual retrieval, training the concept-detector with the factorized representation may also increase the QBSE performance, with the benefit of lower dimensional vectors and reduced training time.

4.5 Conclusions and future work

The paper presented a strategy for video indexing, which combines different information sources in a common latent semantic representation. The experimental evaluation suggests that involving complementary data modalities improve the representation of the video content with the corresponding positive impact in retrieval performance. Multimodal latent semantic indexing strategies has been previously used to tackle content-based image retrieval problems. However, up to our knowledge, this is the first attempt to use this approach for human action video retrieval.

For the particular task addressed in this paper, human action retrieval, the results show that action attributes provide effective cues for action retrieval. By using complementary or finer-grained descriptions of actions (i.e. attributes) along with the action labels, the latent space generated by the method is improved. This results are encouraging and allow us to hypothesize that involving an additional data modality, such as audio, may improve significantly the index quality. However, in all the cases, QBSE outperformed the different multimodal latent semantic indexing strategies. The main reason is the supervised discriminative nature of QBSE which takes advantage of the way relevance is assessed in the experimental evaluation. We still believe the multimodal latent semantic indexing approach has advantages such as its scalability in terms of the size of both the number of images and the vocabulary used to tag them, however, this needs to be evaluated in a more realistic setup that includes large image collections with noisy natural language annotations. This is part of our future work.

Learning from large-scale data is possible through the use of online methods which reduce the memory complexity of the algorithm. Future experimentation will focus on evaluating the scalability of the method using larger video collections.

As the index built is multimodal, queries can be performed not only with visual content but also with the other modalities or a combination of them. This is one path to be further explored to asses the quality of the index. It is also possible to add noise to the queries to evaluate the robustness of the method.

Acknowledgements

This work was partially funded by the project Multimodal Image Retrieval to Support Medical Case-Based Scientific Literature Search, ID R1212LAC006 by Microsoft Research LAC-CIR and Colciencias grant 617 “Jóvenes Investigadores 2013”.

5 Annotating and Retrieving Human Actions using Matrix Factorization

This paper presents a method for annotating and retrieving videos of human actions based on two-way matrix factorization. The method addresses the problem by modeling it as the problem of finding common latent space representation for multimodal objects. In this particular case, the modalities correspond to the visual and textual (annotations) information associated with videos, which are projected by the method to the latent space. Assuming this space exists, it is possible to map between input spaces, i.e. visual to textual, by projecting across the latent space. The mapping between the spaces is explicitly optimized in the cost function and learned from training data including both modalities. The algorithm may be used for annotation, by projecting only visual information and obtaining a textual representation, or for retrieval by indexing on the latent or textual spaces. Experimental evaluation shows competitive results when compared to state-of-the-art annotation and retrieval methods.

This chapter presents the work submitted to "Congreso Iberoamericano de Reconocimiento de Patrones 2015", and is currently under review for acceptance. It is joint work with PhD Fabio González.

5.1 Introduction

Video analysis is a task that has gained a lot of attention by the research community due to the growing amount of video content on the web. Having such large video repositories, e.g. YouTube, poses the problem of efficiently managing and accessing them. A system able to correctly annotate videos is useful for searching, categorizing and understanding applications. This work focuses on video annotation with a predefined number of human actions, according to their presence in the video. For this, only visual information is available at test time, but more information sources may be used during model training, such as action annotations or related audio.

For the current experimentation, the focus is given to the annotation model. The representation problem is not approached, and a state-of-the-art representation is used. The chosen model is two way matrix factorization [63], which has been also used for image annotation. The basic idea of the algorithm is to obtain a common latent space, which embeds the information content of the different modalities. Every modality has two projection matrices, one that allows to map from the original modality space to the latent space, and another

one which maps back from the latent space to the original modality space. By means of this matrices, one could project from a desired input space to another modal space, e.g. from visual to textual by first mapping from visual to latent and finally from latent to textual.
After training the model for the annotation task, we also evaluate it in an information retrieval setup using the same dataset. For retrieval, the query-by-example method is used so only visual information is available as input query. The approach maps the input query visual representation to one of the available spaces, e.g. textual, latent, visual, and the collection videos to be retrieved are also mapped to that space. Having both query and collection in the same representation space, a particular similarity measure could be used to rank videos according to their similarity with the query. The collection videos are then presented to the user according to the ranking. The results show a competitive performance in both tasks when compared to an annotation model and a query by semantic example retrieval mode, both based on linear support vector machines (SVM).
This document is organized as follows: first, a short review of previous work is given on Sec. 5.2; then, the selected annotation model is described in Sec. 5.3; finally, experimental results are presented in Sec. 5.4 and conclusions in Sec. 5.5.

5.2 Previous Work

The evaluation in [68] compares different types of video features and interest point detectors using Bag of Features (BoF) and nonlinear SVM. The most important finding of the evaluation was that dense sampling improved action classification over interest point detectors, with the drawback of large number of features to process. This motivated the proposal of Dense Trajectory Features (DTF) [66] and Improved Dense Trajectories (IDT) [67]. The main idea of these features is to extract four descriptors along volumes aligned with tracked pixels. Tracking is based on dense optical flow with median filtering to determine the new coordinates of a pixel. The four descriptors are

1. Trajectories, a stacking of the X and Y difference between consecutive pixel coordinates. Using these features do not improve performance significantly so it is not commonly used

2. Histogram of Gradients (HoG), which encodes information of appearance

3. Histogram of Flow (HoF), which encodes motion information

4. Motion Boundary Histogram (MBH), generated from the derivative of flow and useful to reduce camera motion

These features have achieved state-of-the-art performance on various action recognition datasets. The IDT proposal used Fisher vectors [23] instead of BoF to encode the features and generate a video representation. Fisher vectors generate a high dimensional representation based on the gradient of local features with respect to the parameters of a generative model of the features.

By the time trajectory features were proposed, Convolutional Neural Networks (CNN) [35] were causing a revolution on image representation. These networks benefited from training deep models with large amounts of data, which allowed to learn complex features and avoid overfitting. Inspired by this trend, many video representations based on CNNs have been proposed. Among them is the work of Simonyan et al. [56], were an appearance and a motion CNN are combined to generate features which are aggregated with average pooling to generate a video representation, achieving a competitive result over IDT. Instead of average or max-pooling, Xu et al. [71] propose to use Fisher vectors to encode CNN features. This proposal achieved excellent results at the event detection task by means of appearance information only.

As IDT have the drawback of large number of generated features and long processing times, Motion Flow (MF) was proposed in [31], which reduces drastically the extraction time by exploiting the flow generated by the MPEG4 video encoding. This features are fast to compute, but achieve a lower performance compared to IDT.

In [68] and [66], a nonlinear SVM was used to assign single actions to a previously segmented video, represented using BoF. The nonlinear kernel in those cases was χ^2. Where multiple descriptors were available, such as for IDT, the kernels were aggregated by addition before applying the exponential function.

Representations with Fisher vectors generate high dimensional vectors, which achieve good performance with linear models, and so linear SVMs are used in [67] and [71].

5.3 Annotation Model

Two Way Matrix Factorization (TWMF) is a latent space method in which the cost function to optimize has no explicit mention of the latent space, and the objective is to obtain the most accurate reconstruction of the modalities when projected through the latent space and back. The basic assumption is that there exists a linear projection matrix between two information modalities, e.g. textual and visual. This projection first maps one modality to the latent space, and then projects from that space to the other modality. Lets assume we have vectorial representations $v_i \in \mathbb{R}^{D_v}$ and $t_i \in \mathbb{R}^{D_t}$ of the two modalities (v, t) for a given entity i. Each of the modalities has a projection matrix to the latent space, $W_t \in \mathbb{R}^{r \times D_t}$ and $W_v \in \mathbb{R}^{r \times D_v}$ respectively, where r is the dimension of the latent space. After projecting, a latent representation $h_i \in \mathbb{R}^{r \times 1}$ is obtained for the entity.

$$h_i = W_t t_i \ . \tag{5.1}$$

$$h_i = W_v v_i \ . \tag{5.2}$$

Each of the modalities has also a back-projection matrix, $W_t' \in \mathbb{R}^{D_t \times r}$ and $W_v' \in \mathbb{R}^{D_t \times r}$, which maps back from the latent space to the respective modality.

$$\begin{aligned} t_i &= W_t' h_i \ . \qquad (5.3) \\ v_i &= W_v' h_i \ . \qquad (5.4) \end{aligned}$$

To express the relationship between modalities we can combine the previous expressions to obtain crossmodal mappings that don't explicitly use the latent space:

$$\begin{aligned} t_i &= W_t' W_v v_i \ . \qquad (5.5) \\ v_i &= W_v' W_t t_i \ . \qquad (5.6) \end{aligned}$$

Assuming we have both representations for a number n of entities, the data for a modality can be conveniently represented in a matrix were each column represents an entity. Following our naming convention, these matrices are called $V \in \mathbb{R}^{D_v \times n}$ and $T \in \mathbb{R}^{D_t \times n}$. In the case of projecting from the "visual" to the "textual" modality, we would have

$$T = W_t' W_v V \ . \qquad (5.7)$$

To obtain the projection and back-projection matrices we use Stochastic Gradient Descent implemented in Pylearn2 [17]. A cost function is minimized by letting the library calculate gradients. For the annotation task, the interest is in assigning labels based on visual content. The main term of the cost is based on least squares reconstruction of the textual modality from the projection through the latent space of the visual modality

$$\arg\min_{W_t', W_v} ||T - W_t' W_v V||_F^2 \ . \qquad (5.8)$$

What this means is that we want a model that reconstructs as accurately as possible the textual modality from the visual modality.

Additional terms can be added to control model complexity, overfitting and sparsity among others. The initial formulation in [63] adds a regularization term of the projection matrices and reconstruction of each modality after passing through the latent space and back (one way), resulting in the following optimization problem:

$$\arg \min_{W_t, W_t', W_v W_v'} \begin{pmatrix} \delta ||T - W_t' W_v V||_F^2 + \\ \alpha ||T - W_t' W_t T||_F^2 + \\ (1-\alpha) ||V - W_v' W_v V||_F^2 + \\ \beta \left(||W_t'||_F^2 + ||W_v||_F^2 + ||W_t||_F^2 + ||W_v'||_F^2 \right) \end{pmatrix} \ . \qquad (5.9)$$

5.4 Experiments

5.4.1 Annotation task and dataset description

The goal of the task is to recognize the action or actions present in a video from a predefined set of actions. For the experimental evaluation we used two publicly available datasets: UCF101 [61] and THUMOS 2014 [25]. The UCF101 dataset contains realistic videos where each clip has exactly one action and has been segmented in time to have the best fit to the action. The dataset comprises 101 different actions, each action with several example videos. The THUMOS 2014 dataset includes videos belonging to the same 101 actions, but without time segmentation, so a single video can have more than one execution of multiple activities and also frames from activities different to the 101 actions set. There are two tasks and the current evaluation is performed on the first task: recognition. The second task is temporal segmentation.

The dataset has a training partition of 13320 trimmed videos from UCF101 dataset, a validation partition of 1010 untrimmed videos from THUMOS 2014 and a test partition of 1574 untrimmed videos also from THUMOS 2014.

Validation data may be used as part of training data to generate test results. For each test video, the objective is to generate a score for each of the 101 actions present in the training set. In the case of untrimmed videos, more than one action may be present in each sequence. Results are evaluated using Mean Average Precision (MAP).

Two models are evaluated: linear SVMs and TWMF. Models are trained using only training data, and evaluated on validation and test data. We also train the models by augmenting training data with validation data and evaluate them on test data only.

In the THUMOS data, a score must be generated for each action. For TWMF, the textual projection value is used as score and for linear SVM, Platt scaled scores. The SVM implementation is from Scikit-Learn [46].

Visual and textual representation

As visual features we use IDT, extracted using the implementation provided by the IDT author and encoded using Fisher vectors using the pipeline described in [47]. The resulting Fisher Vectors are l^2 and power normalized as it is a recommended practice. The Fisher Vector implementation used is VlFeat [65]. The size of the resulting visual representation using Fisher Vectors when concatenating the 4 descriptors is 101,376.

The textual information consists of a binary 101-dimensional vector. Each dimension represents an action, and a 1 value indicates presence, while a 0 indicates absence, of the action.

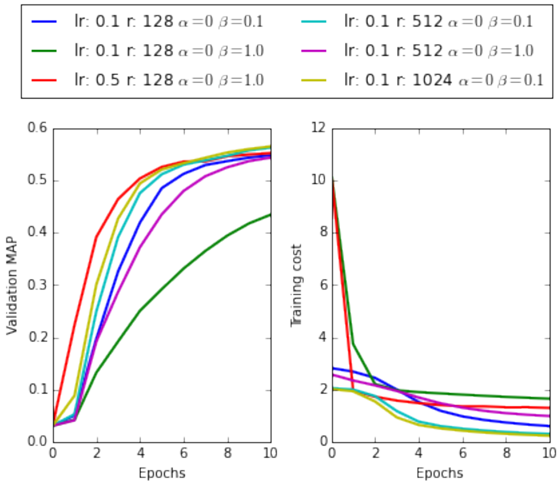


Figure 5.1: Validation MAP and training cost when using all reconstruction terms and regularization of all projection matrices

Table 5.1: MAP comparison between linear SVM and TWMF.

Method	Validation	Test
Linear SVM C=100	52.6%	39.8%
two-way MF	57.2%	42.4%

Two Way Matrix Factorization Cost Components Evaluation

To select the algorithm parameters, we performed experiments using only parts of the cost function and evaluating its influence on the validation set MAP and training set cost value during each epoch. The most important term after the two way reconstruction is the one way reconstruction which speeds up convergence. Giving high importance to the regularization slows convergence. The results when the complete cost is used are presented in Fig. 5.1. It includes one experiment with a learning rate of 0.5. Other experiments using that learning rate showed a faster convergence rate but instability problems when the validation MAP was saturated. The specific case in the figure seem to have converged due to the high regularization. Only experiments with $\alpha = 0$ are presented as giving importance to reconstructing the one way visual term slowed convergence and in some cases produced instabilities.

Results

The MAP results for linear SVM and TWMF are presented in Tab. 5.1. TWMF performs slightly better than linear SVM. Both algorithms show a drop of approximately 10% on test data with respect to validation data.

The Two Way algorithm was trained using GPU acceleration provided by Theano and Pylearn2. Two GPU devices were available: a Tesla C2050 and a K40. The first one takes

Table 5.2: Comparison of retrieval results for visual retrieval, QBSE and TWMF using the latent and textual spaces.

Method	Validation	Test
Visual	11.8%	9.4%
QBSE SVM	51.5%	39.6%
Two Way MF Textual 10 epochs	52.5%	39.5%
Two Way MF Latent 10 epochs	52.5%	39.4%
Two Way MF Textual 20 epochs	57.9%	43.5%
Two Way MF Latent 20 epochs	58.4%	43.6%

approximately 30 seconds per epoch for training and the latter takes around 12 seconds which is almost a speed up of 3 times. Training the SVM including the kernel calculation takes almost 8 minutes. Using GPU acceleration, half of the time to train a linear SVM is needed to train a two-way model for 20 epochs. When no GPU acceleration is used, each epoch takes a minute in training. In the case of the SVM, the kernel calculation is parallelized in all the available CPU cores.

5.4.2 Retrieval experiments

After evaluating both methods on the recognition task, we are interested on the retrieval task. The setup for the retrieval experiments consists on taking a partition of the dataset as the collection used by the system to retrieve documents, and the remaining documents as queries. In this case query-by-example is used, so each query visual representation is used as input to the system, and the aim is that the system generates a ranked list of relevant documents. Documents are considered relevant if they share at least one of the actions in the query document. To generate the ranked results, both query and collection are mapped to a semantic space. Using a similarity measure, the semantic representation of the query is compared to the semantic representation of each of the documents in the collection. The similarity value is used to generate the ranking. The semantic spaces considered are the textual space, corresponding to the action labels, and the latent space. Following previous work, we compare retrieval using TWMF with pure visual retrieval and QBSE using the learned linear SVM. The similarity measure is dot product in all cases, and the evaluation metric is also Mean Average Precision (MAP). The results are presented in Tab. 5.2.

5.5 Conclusions

The TWMF algorithm is an efficient method for annotation and retrieval of videos containing actions. Its online nature allows to train on large datasets without huge memory requirements and achieving task performance competitive or even better than using linear SVM. By means of GPU acceleration, the algorithm also achieves shorter training times compared to the linear SVM implementation which is also accelerated but with CPU parallelization of the kernel calculation.

The state of the art on the THUMOS dataset is achieved by combining IDT features with CNN features. Including CNN features as an additional modality or simply concatenating them with IDT is part of the future work. Nevertheless, the linear SVM baseline with IDT features is a good enough baseline when the comparison is focused not on the features but on the learning model.

6 Conclusions and future work

6.1 Conclusions

The selected features play an important role on the performance of the algorithm. Along our research, we have evaluated 3 types of features: HOG/HOF extracted according to space-time interest point detectors, Dense Trajectory Features focusing on the MBH descriptor which is the best performing descriptor by itself, and finally Improved Trajectory Features using the HOG, HOF and MBH descriptors. Part of the improvement in performance achieved is attributed to improving the features. The improved features have come along with an important drawback: long extraction times and increased storage requirements. This drawback may be mitigated by encoding the features online instead of waiting for the full extraction process to terminate. This way only the encoded representation has to be stored instead of the full set of features, so a constant storage for a fixed length vector is required per video sequence and not a varying number of fixed length descriptors.

Obtaining video features that are at the same time fast to extract and descriptive enough is still an open problem. Though this work only applied online encoding, some other authors have also proposed sampling the densely extracted features. This approach has also a drawback which is loss in performance. CNN based features are a promising alternative which is receiving a lot of attention and have achieved similar performance as IDT. An advantage of these features is that they can be speed up by means of GPU devices.

Despite the gains obtained by using CNN proposals, we consider IDT as a good enough feature as our focus is on the machine learning method instead of the features. But for a practical retrieval system, more efficient features must be developed which allow almost real time processing to avoid degrading the user experience.

As in the case of features, human action datasets have also evolved drastically in the last years. More realistic videos are used and also the number of samples and categories has increased. One of the drawbacks of the used datasets is the limited number of action classes which limits the diversity of the possible queries to the system. That limitation is partially addressed by using a query by example approach, which allows to present diverse queries. Nevertheless, relevance is still judged according to action labels and that is a limitation hard to overcome as there is little or no additional metadata that may be used to determine relevance. Recent work has made available new datasets aimed at increasing the number of samples and action classes, applying the proposed methods to those dataset may be helpful in verifying the effectiveness of the algorithms.

Along the research the proposed methods have achieved improvements in efficiency and task performance. This improvements can be attributed to the employment of online training instead of batch processing of data and to the objective formulation switch from unsupervised to supervised. As we have previously mentioned, part of the improvement is due also to the improved features and encoding, but those are not components of the learning method.

The efficiency improvement is evidenced by the ability to process a larger number of samples (6,676 video sequences in UCF50 versus 13,320 in UCF101) with higher dimensional representations (4,000 dimensions with BoF encoding versus 101,376 with Fisher Vector encoding). First of all, the online algorithm reduces the memory requirements for a learning iteration which allows to train the algorithm even on limited memory machines such as a non cutting edge personal laptop or desktop computers. To achieve efficiency we also leveraged tools like the HDF5 format which also reduces the memory foot print and by means of compression reduces disk storage requirements and allows a hierarchical organization of data. Finally, as matrix factorization is an algorithm with intensive use of dot products, CPU and GPU acceleration of that operation was possible by means of tools like Intel's Math Kernel Library for multicore CPU acceleration and Theano/Pylearn2 for GPU acceleration.

The retrieval performance was also improved by using a supervised objective formulation. As the work presented in Chapter 4 made evident, an algorithm as QBSE backed up by a supervised method like SVMs outperformed our matrix factorization proposal which was based on an unsupervised objective function aimed at reconstruction of the input matrices. By including explicitly the reconstruction of textual data from visual data in the objective, the work presented in Chapter 5 achieved a competitive performance with respect to SVMs and in some cases was able to improve it. The comparison was made using the same features for both methods. Those results confirm the validity of the proposed method to tackle the annotation and retrieval problems.

The retrieval experiments using the textual projection is a similar proposal as QBSE but based on matrix factorization, as both methods focus on a semantic space used for retrieval obtained in a supervised manner. One feature of the proposed factorization is that a latent representation is obtained that improves retrieval even though the latent factors are learned by the algorithm.

6.2 Future work

Using some of the most recent datasets is a path to explore. This would allow us to explore retrieval in a more realistic setting where the possible user information needs are broader. But to start working in this direction, first we must change the features employed as IDT are computationally expensive and would represent a bottleneck when handling larger scale datasets. The first alternatives in that direction are CNN based proposals as the research group has access to a Tesla K40 GPU donated by Nvidia.

An additional path to explore is the generation of synthetic textual descriptions as an additional modality of data. One example method could be based on Recurrent Neural Networks (RNN) [32]. This synthetic descriptions, though not perfect, might bring a wider textual vocabulary and allow us to build a system with more semantic concepts to be used on queries.

Bibliography

[1] Andrade, Felipe S P. ; Almeida, Jurandy: Fusion of Local and Global Descriptors for Content-Based Image and Video Retrieval. (2012), S. 845–853

[2] Belkhatir, Mohammed ; Alhashmp, Saadat M. ; Daniep, Thomas O.: Combining Content and Context Information for Video Events Classification and Retrieval. , S. 81–86

[3] Berners-Lee, Tim: The Semantic Web. (2001), Nr. May, S. 1–4

[4] Bottou, Léon: Large-Scale Machine Learning with Stochastic Gradient Descent. In: Lechevallier, Yves (Hrsg.) ; Saporta, Gilbert (Hrsg.): *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*. Paris, France : Springer, August 2010, S. 177–187

[5] Caicedo, Juan C. ; BenAbdallah, Jaafar ; González, Fabio A. ; Nasraoui, Olfa: Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. In: *Neurocomputing* 76 (2012), Nr. 1, S. 50–60

[6] Caicedo, Juan C. ; González, Fabio A.: Multimodal fusion for image retrieval using matrix factorization. In: *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval*. New York, NY, USA : ACM, 2012 (ICMR '12). – ISBN 978–1–4503–1329–2, S. 56:1–56:8

[7] Caicedo, Juan C. ; González, Fabio A.: Online Matrix Factorization for Multimodal Image Retrieval. In: Alvarez, Luis (Hrsg.) ; Mejail, Marta (Hrsg.) ; Gomez, Luis (Hrsg.) ; Jacobo, Julio (Hrsg.): *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* Bd. 7441. Springer Berlin Heidelberg, 2012. – ISBN 978–3–642–33274–6, S. 340–347

[8] Chaquet, Jose M. ; Carmona, Enrique J. ; Fernández-Caballero, Antonio: A Survey of Video Datasets for Human Action and Activity Recognition. In: *Computer Vision and Image Understanding* 117 (2013), Februar, Nr. 6, S. 633–659. – ISSN 10773142

[9] Dementhon, Daniel ; Doermann, David: Video Retrieval using Spatio-Temporal Descriptors. (2003). ISBN 1581137222

[10] Ding, Chris H. Q. ; Li, Tao ; Peng, Wei: On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing. In: *Computational Statistics and Data Analysis* 52 (2008), Nr. 8, S. 3913–3927

[11] Dollár, Piotr ; Rabaud, Vincent ; Cottrell, Garrison ; Belongie, Serge: Behavior recognition via sparse spatio-temporal features. In: *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on* IEEE, 2005, S. 65–72

[12] Ericsson: Ericsson Mobility Report. (2013), Nr. June

[13] Eweiwi, Abdalrahman ; Cheema, Muhammad S. ; Bauckhage, Christian: Discriminative Joint Non-negative Matrix Factorization for Human Action Classification. In: *Pattern Recognition.* Springer, 2013, S. 61–70

[14] Fabian Caba Heilbron, Bernard G. ; Niebles, Juan C.: ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, S. 961–970

[15] Forsyth, D.A. ; Ponce, J.: *Computer Vision: A Modern Approach.* Pearson Education, Limited, 2011 (Always learning). – ISBN 9780136085928

[16] Forsyth, David ; Ponce, Jean: Computer Vision, A Modern Approach. (2003)

[17] Goodfellow, Ian J. ; Warde-Farley, David ; Lamblin, Pascal ; Dumoulin, Vincent ; Mirza, Mehdi ; Pascanu, Razvan ; Bergstra, James ; Bastien, Frédéric ; Bengio, Yoshua: Pylearn2: a machine learning research library. In: *arXiv preprint arXiv:1308.4214* (2013)

[18] Hassner, Tal: A Critical Review of Action Recognition Benchmarks. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on* IEEE, 2013, S. 245–250

[19] Hauptmann, Alexander ; Yan, Rong ; Lin, Wei-hao ; Christel, Michael ; Wactlar, Howard: Filling the Semantic Gap in Video Retrieval: An Exploration. (2008), S. 253–278

[20] Healey, RG: Database management systems. In: *Geographic Information Systems: Principles and Applications* 1 (1991), S. 251–267

[21] Hu, Weiming ; Xie, Dan ; Fu, Zhouyu ; Zeng, Wenrong ; Maybank, Steve: Semantic-based surveillance video retrieval. In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 16 (2007), April, Nr. 4, S. 1168–81. – ISSN 1057–7149

[22] Ikizler, Nazli ; Forsyth, David: Searching Video for Complex Activities with Finite State Models. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition* (2007), Juni, S. 1–8. ISBN 1–4244–1179–3

[23] Jaakkola, Tommi ; Haussler, David [u. a.]: Exploiting generative models in discriminative classifiers. In: *Advances in neural information processing systems* (1999), S. 487–493

[24] Jiang, Y.-G. ; Liu, J. ; Roshan Zamir, A. ; Laptev, I. ; Piccardi, M. ; Shah, M. ; Sukthankar, R. *THUMOS Challenge: Action Recognition with a Large Number of Classes.* 2013

[25] Jiang, Y.-G. ; Liu, J. ; Roshan Zamir, A. ; Toderici, G. ; Laptev, I. ; Shah, M. ; Sukthankar, R. *THUMOS Challenge: Action Recognition with a Large Number of Classes.* `http://crcv.ucf.edu/THUMOS14/`. 2014

[26] Jones, Simon ; Shao, Ling: Action retrieval with relevance feedback on YouTube videos. In: *Proceedings of the Third International Conference on Internet Multimedia Computing and Service - ICIMCS '11* (2011), S. 42. ISBN 9781450309189

[27] Jones, Simon ; Shao, Ling: Action retrieval with relevance feedback on youtube videos. In: *Proceedings of the Third International Conference on Internet Multimedia Computing and Service* ACM, 2011, S. 42–45

[28] Jones, Simon ; Shao, Ling: Content-Based Retrieval of Human Actions from Realistic Video Databases. In: *Information Sciences* 236 (2013), Februar, S. 56–65. – ISSN 00200255

[29] Jones, Simon ; Shao, Ling ; Zhang, Jianguo ; Liu, Yan: Relevance feedback for real-world human action retrieval. In: *Pattern Recognition Letters* 33 (2012), März, Nr. 4, S. 446–452. – ISSN 01678655

[30] Jones, Simon ; Shao, Ling ; Zhang, Jianguo ; Liu, Yan: Relevance feedback for real-world human action retrieval. In: *Pattern Recognition Letters* 33 (2012), Nr. 4, S. 446–452

[31] Kantorov, Vadim ; Laptev, Ivan: Efficient Feature Extraction, Encoding, and Classification for Action Recognition. In: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* IEEE, 2014, S. 2593–2600

[32] Karpathy, Andrej ; Fei-Fei, Li: Deep visual-semantic alignments for generating image descriptions. In: *arXiv preprint arXiv:1412.2306* (2014)

[33] Karpathy, Andrej ; Toderici, George ; Shetty, Sanketh ; Leung, Thomas ; Sukthankar, Rahul ; Fei-Fei, Li: Large-scale video classification with convolutional neural networks. In: *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on* IEEE, 2014, S. 1725–1732

[34] Kliper-Gross, Orit ; Gurovich, Yaron ; Hassner, Tal ; Wolf, Lior: Motion interchange patterns for action recognition in unconstrained videos. In: *Proceedings of the 12th European conference on Computer Vision - Volume Part VI*. Berlin, Heidelberg : Springer-Verlag, 2012 (ECCV'12). – ISBN 978–3–642–33782–6, S. 256–269

[35] Krizhevsky, Alex ; Sutskever, Ilya ; Hinton, Geoffrey E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, 2012, S. 1097–1105

[36] Kuehne, H. ; Jhuang, H. ; Garrote, E. ; Poggio, T. ; Serre, T.: HMDB: A large video database for human motion recognition. In: *2011 International Conference on Computer Vision* (2011), November, S. 2556–2563. ISBN 978–1–4577–1102–2

[37] Lamard, Mathieu ; Cazuguel, Guy ; Droueche, Zakarya ; Roux, Christian: Real-Time Retrieval of Similar Videos with Application to Computer-Aided Retinal Surgery. (2011), S. 4465–4468. ISBN 9781424441228

[38] Laptev, I. ; Marszalek, M. ; Schmid, C. ; Rozenfeld, B.: Learning realistic human actions from movies. In: *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008. – ISSN 1063–6919, S. 1–8

[39] Laptev, Ivan: On Space-Time Interest Points. In: *Int. J. Comput. Vision* 64 (2005), September, Nr. 2-3, S. 107–123. – ISSN 0920–5691

[40] Lee, Daniel D. ; Seung, H S.: Learning the parts of objects by non-negative matrix factorization. In: *Nature* 401 (1999), Nr. 6755, S. 788–791

[41] Lee, Daniel D. ; Seung, H S.: Algorithms for non-negative matrix factorization. In: *Advances in neural information processing systems*, 2000, S. 556–562

[42] Manning, C. ; Raghavan, P ; Schütze, H.: An Introduction to Information Retrieval. (2009), Nr. c

[43] Niebles, Juan C. ; Wang, Hongcheng ; Fei-Fei, Li: Unsupervised learning of human action categories using spatial-temporal words. In: *International journal of computer vision* 79 (2008), Nr. 3, S. 299–318

[44] Ning, Huazhong ; Hu, Yuxiao ; Huang, Thomas S. ; Avenue, North M.: SEARCHING HUMAN BEHAVIORS USING SPATIAL-TEMPORAL WORDS University of Illinois at Urbana-Champaign.

[45] Otálora-Montenegro, Sebastian ; Pérez-Rubiano, Santiago A. ; González, Fabio A.: Online Matrix Factorization for Space Embedding Multilabel Annotation. In: *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications.* Springer, 2013, S. 343–350

[46] Pedregosa, F. ; Varoquaux, G. ; Gramfort, A. ; Michel, V. ; Thirion, B. ; Grisel, O. ; Blondel, M. ; Prettenhofer, P. ; Weiss, R. ; Dubourg, V. ; Vanderplas, J. ; Passos, A. ; Cournapeau, D. ; Brucher, M. ; Perrot, M. ; Duchesnay, E.: Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12 (2011), S. 2825–2830

[47] Peng, Xiaojiang ; Wang, Limin ; Wang, Xingxing ; Qiao, Yu: Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. In: *arXiv preprint arXiv:1405.4506* (2014)

[48] Rasiwasia, Nikhil ; Moreno, Pedro J. ; Vasconcelos, Nuno: Bridging the gap: Query by semantic example. In: *Multimedia, IEEE Transactions on* 9 (2007), Nr. 5, S. 923–938

[49] Reddy, Kishore K. ; Shah, Mubarak: Recognizing 50 human action categories of web videos. In: *Machine Vision and Applications* (2012), November. – ISSN 0932–8092

[50] Rubner, Yossi ; Tomasi, Carlo ; Guibas, Leonidas J.: The earth mover's distance as a metric for image retrieval. In: *International Journal of Computer Vision* 40 (2000), Nr. 2, S. 99–121

[51] Ruiz, Pablo ; Babacan, S D. ; Molina, Rafael ; Katsaggelos, Aggelos K.: Retrieval of Video Clips with Missing Frames using Sparse Bayesian Reconstruction . (2011), Nr. Ispa, S. 443–448

[52] Shah, M.: Recognizing realistic actions from videos "in the wild". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), Juni, S. 1996–2003. ISBN 978–1–4244–3992–8

[53] Shi, Jianbo ; Tomasi, Carlo: Good features to track. In: *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on* IEEE, 1994, S. 593–600

[54] Sicre, R ; Nicolas, H: Human Behavior Analysis at a Point of Sale. (2010), S. 635–644

[55] Sicre, R ; Nicolas, H: Human behavior analysis at a point of sale. In: *Advances in Visual Computing.* Springer, 2010, S. 635–644

[56] Simonyan, Karen ; Zisserman, Andrew: Two-stream convolutional networks for action recognition in videos. In: *Advances in Neural Information Processing Systems*, 2014, S. 568–576

[57] Sivic, Josef ; Zisserman, Andrew: Efficient Visual Search of Videos Cast as Text Retrieval. 31 (2009), Nr. 4, S. 591–606

[58] Smith, John R. ; Naphade, Milind ; Natsev, Apostol: Multimedia semantic indexing using model vectors. In: *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on* Bd. 2 IEEE, 2003, S. II–445

[59] Snoek, Cees G. ; Worring, Marcel ; Smeulders, Arnold W.: Early versus late fusion in semantic video analysis. In: *Proceedings of the 13th annual ACM international conference on Multimedia* ACM, 2005, S. 399–402

[60] Soomro, K. ; Roshan Zamir, A. ; Shah, M.: UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. In: *CRCV-TR-12-01*, 2012

[61] Soomro, Khurram ; Zamir, Amir R. ; Shah, Mubarak: Ucf101: A dataset of 101 human actions classes from videos in the wild. In: *arXiv preprint arXiv:1212.0402* (2012)

[62] Vallet, David ; Cantador, Iván ; Jose, Joemon M.: Exploiting semantics on external resources to gather visual examples for video retrieval. (2012). ISBN 1373501200

[63] Vanegas, Jorge A. ; Beltran, Viviana ; González, Fabio A.: Two-way Multimodal Online Matrix Factorization for Multi-label Annotation. In: *International Conference on Pattern Recognition Applications and Methods*, 2015, S. 279–285

[64] Vanegas, Jorge A. ; Caicedo, Juan C. ; González, Fabio A. ; Romero, Eduardo: Histology Image Indexing Using a Non-negative Semantic Embedding. In: Müller, Henning (Hrsg.) ; Greenspan, Hayit (Hrsg.) ; Syeda-Mahmood, Tanveer (Hrsg.): *Medical Content-Based Retrieval for Clinical Decision Support* Bd. 7075. Springer Berlin Heidelberg, 2012. – ISBN 978–3–642–28459–5, S. 80–91

[65] Vedaldi, A. ; Fulkerson, B. *VLFeat: An Open and Portable Library of Computer Vision Algorithms*. `http://www.vlfeat.org/`. 2008

[66] Wang, Heng ; Klaser, Alexander ; Schmid, Cordelia ; Liu, Cheng-Lin: Action recognition by dense trajectories. In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on* IEEE, 2011, S. 3169–3176

[67] Wang, Heng ; Schmid, Cordelia: Action recognition with improved trajectories. In: *Computer Vision (ICCV), 2013 IEEE International Conference on* IEEE, 2013, S. 3551–3558

[68] Wang, Heng ; Ullah, Muhammad M. ; Klaser, Alexander ; Laptev, Ivan ; Schmid, Cordelia: Evaluation of local spatio-temporal features for action recognition. In: *BMVC 2009-British Machine Vision Conference* BMVA Press, 2009, S. 124–1

[69] Wang, Lei ; Song, Dawei ; Elyan, Eyad: LNCS 6611 - Video Retrieval Based on Words-of-Interest Selection. (2011), S. 687–690

[70] Xiong, Ziyou ; Zhou, Xiang S. ; Tian, Qi ; Rui, Yong ; Huang, Thomas S.: Semantic Retrieval of Video [. (2006), Nr. March, S. 18–27

[71] Xu, Zhongwen ; Yang, Yi ; Hauptmann, Alexander G.: A Discriminative CNN Video Representation for Event Detection. In: *arXiv preprint arXiv:1411.4006* (2014)