# A systemic approach to automatic metadata extraction from multimedia content

Christos Varytimidis, Georgios Tsatiris, Konstantinos Rapantzikos
*School of Electrical and Computer Engineering*
*National Technical University of Athens*
*9, Heroon Politechniou str., 15780*
*Athens, Greece.*
*Email: [chrisvar,gtsatiris,rap]@image.ntua.gr*

Stefanos Kollias
*School of Computer Science*
*University of Lincoln*
*Brayford Pool, Lincoln*
*Lincolnshire, UK.*
*Email: skollias@lincoln.ac.uk*

***Abstract*—There is a need for automatic processing and extracting of meaningful metadata from multimedia information, especially in the audiovisual industry. This higher level information is used in a variety of practices, such as enriching multimedia content with external links, clickable objects and useful related information in general. This paper presents a system for efficient multimedia content analysis and automatic annotation within a multimedia processing and publishing framework. This system is comprised of three modules: the first provides detection of faces and recognition of known persons; the second provides generic object detection, based on a deep convolutional neural network topology; the third provides automated location estimation and landmark recognition based on state-of-the-art technologies. The results are exported in meaningful metadata that can be utilized in various ways. The system has been developed and successfully tested in the framework of the EC Horizon 2020 Mecanex project, targeting advertising and production markets.**

***Index Terms*—object detection, face recognition, landmark recognition, deep neural networks, bag-of-words.**

## 1. Introduction

One of the major challenges for enterprises engaged in the arena of multimedia content creation and usage (e.g. production and post-production companies, advertising agencies, online publishing companies, etc) is the development of business models to use the advantages of modern devices and generate new revenue streams from the growing audience on new - mobile - multimedia systems. In such a typical scenario, multimedia content after creation, is enriched and expanded with new content, as well as enabled to provide access to relevant online material. Providing enterprises involved in production and post-production with tools to initially enrich their multimedia content with relevant information can enhance their capabilities and sources of revenue streams.

### 1.1. The proposed system at a glance

Motivated by the aforementioned trends, the aim of the project[1] in which our system has firstly been developed has been to provide innovative tools and methods that enable the automatic annotation and editorial process during the creation, production and post-production of multimedia content that will allow enterprises to create and enrich audiovisual data ready for multi-screen environments and marketing campaigns. It also supports search and retrieval mechanisms from libraries with existing enriched multimedia content that could be used as building blocks of a fast innovative creative process.

Another direction of the project has been to pair created metadata with multimedia information allowing easy creation and access to relevant educational and recreational content, e.g. Wikipedia pages of landmarks appearing in a video, images of actors or staff participating, relevant trivia, historic events affecting the story etc. Furthermore, the framework provides a multi-screen toolkit for the development of prototypes that enable an automatic porting to different target platforms, such as regular web pages, mobile pages and mobile apps as well as TV applications.

The final objective of the project has been to gather (from end users or consumers), process and deliver personalized information, which will facilitate the production and editorial process, while at the same time will allow advertising and marketing brokers to automatically ingest relevant ads and marketing data via coupling user preferences with the details of the playing multimedia content. End users will participate within this process providing feedback towards both the enrichment of metadata by manually annotating parts of the content, as well as expressing their views related both to the experience and the content itself.

The system which is presented in this paper has been developed in this framework targeting the automatic annotation of multimedia content, as described next in this section.

1. EC H2020 Horizon "Multimedia Content Annotations for Rapid Exploitation in Multi-Screen Environments" (MECANEX), http://mecanex.eu/, 2014-2016

### 1.2. Automatic annotation of multimedia content

The proposed system processes videos in a per-frame fashion and generates annotations based on the obtained detection and recognition results. Three different image analysis tasks are performed in the automatic annotation pipeline. Section 2 outlines how the task of face detection and recognition of specific people is handled. Section 3 shows the generic object detection process, while section 4 documents how the system is able to identify known landmarks. Experiments and results which show the good performance of the complete pipeline are presented in section 5, followed by conclusions in section 6.

An overview of the system is shown in figure 1. In essence, the input of the system is a video file. The file is broken down into its individual frames and each frame is processed independently. Processing can be done either at certain time intervals or on every frame of the video. A frame is inspected for faces and specific persons, generic object occurrences and depictions of landmarks. After all frames are processed, results are compiled into a single JsON file that is available to the rest of a related platform via a RESTful service.

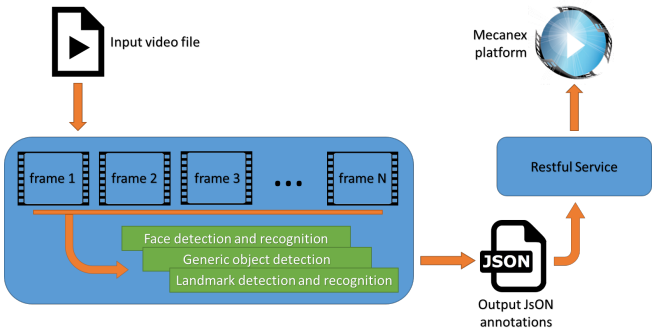



Figure 1. Overview of the automatic annotation system.

## 2. Face detection and recognition

For the task of face detection and recognition, we exploit both well-established computer vision algorithms, as well as more modern approaches that have proven to provide superior performance. We employ a baseline face detector, namely the established Viola and Jones [1] algorithm that provides good detection results even for noisy and low resolution images. For recognition, as a way to assign a label to the detected faces, the Local Binary Patterns Histograms approach of Ahonen et al. [5], which provides state-of-the-art results, was utilized.

Viola and Jones proposed a face detector that is based on a cascade of classifiers, which was later improved in terms of rotation invariance by Lienhart and Maydt [2]. The method uses Haar-like features for describing visual patterns. In the face detection context, a predefined subset of Haar-like features is used, computed by an AdaBoost-based learning algorithm, which selects a small number of critical visual features from a larger set, in order to detect certain visual features in the image that correspond to facial landmarks. Some of these features can be seen in figure 2.

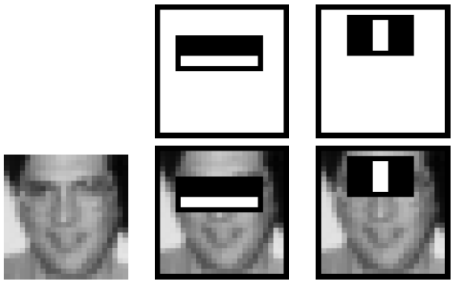


Figure 2. Two Haar-like features. The first feature compares the intensity of the eye and a region across the upper cheeks, while the second feature compares intensities in the eye regions to the intensity over the nose. Figure is taken from [1].

Recognition of known people is a challenging task that is tackled by the presented system. The Eigenfaces method, described in [3], regards a facial image as a point in a high-dimensional image space and aims to compute a lower-dimensional representation, where classification is easier. Principal Component Analysis (PCA) is used to find the lower-dimensional subspace. From a reconstruction point of view, this kind of transformation seems optimal. However, an issue with the Eigenfaces method is that PCA finds a linear projection of features that maximizes the total variance of the data but does not take into account the between-class distances, separating data belonging to different classes. This may lead to loss of class-discriminative information and produce a set of uncorrelated features which have little to offer to the task of face classification. A typical example is when variance between samples is generated from external sources, such as light conditions. So a class-specific projection, using Linear Discriminant Analysis (LDA) instead of PCA, has been used, combined with the Fisherfaces approach [4]. LDA performs class-specific dimensionality reduction by maximizing the ratio of between-classes to within-classes scatter, instead of maximizing the overall scatter.

As both Eigenfaces and Fisherfaces represent faces as points in a high-dimensional space, they require large training sets of facial image samples, which also need to be taken under relatively stable lighting conditions. This isnt always the case, especially in the context of our targeted applications, where classification needs to be performed on video sequences of variable lighting and setup, in cases where large training sets are not always available. In this paper we take into account this requirement and exploit the Local Binary Patterns Histograms approach, proposed by Ahonen et al. [5]. The notion behind LBPH lies in the need to construct a method that can operate with small training datasets (in extreme cases, with a single training image), based on local features robust to translation, rotation and scaling variations. Local Binary Patterns methodology has its roots in two dimnsional texture analysis. The basic idea

is to represent local structure by comparing each pixel with its neighborhood. This is achieved by thresholding every pixel in the neighborhood with the value of the central pixel and constructing an LBP code for that pixel. An example of an early LBP operator is depicted in figure 3. The LBP description captures fine details, but is not scale invariant. In this direction, the operator was extended to use pixel neighborhoods. A circular neighborhood is created around a central pixel with variable radius as scale. Neighbor values are calculated using interpolation when necessary and thus the method represents neighborhoods resembling specific features such as lines, edges and corners.

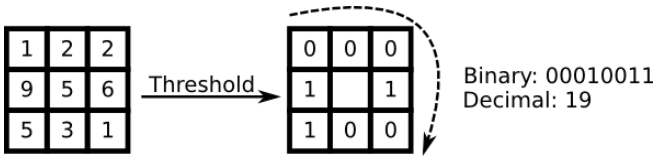



Figure 3. Two Haar-like features. The first feature compares the intensity of the eye and a region across the upper cheeks, while the second feature compares intensities in the eye regions to the intensity over the nose. Figure is taken from [1].

By applying such operators on every possible pixel neighborhood in the image (i.e. using every image pixel as a neighborhood center), a set of LBP codes representing the entire image is constructed. Then, a histogram is calculated, containing information about the distribution of local features such as edges, spots and flat areas, over the whole image. For efficient face representation, since features cannot be located in arbitrary positions, their spatial distribution is also exploited, by dividing an image into sub-regions. The resulting representation offers a description of the face on three different levels of locality: a) the LBP codes contain information about the features on a pixel-level, b) their summation over a region offers information on a regional level, and c) the regional histograms are concatenated to build a global description of the whole image

In the proposed system, we calculate the LBPH representations of all training facial images of famous persons and assign labels to them. When a face is detected in a frame, its LBPH representation is also calculated and compared against the set of training images. The label of the class that is closer in terms of chi square distance ($\tilde{\chi}^2$) to the facial image under inspection is assigned to it. Distance from a class is met with a K-NN approach, to ensure that the labeling is unaffected by spikes.

## 3. Generic object detection

Object recognition is the task of assigning an object category to a whole image, based on its content. This task becomes difficult when images contain many objects, or the main object occupies only a small part of the image, with the rest belonging to background. Thus, we first perform object detection, where the exact position of the object of interest is determined. The output is an object category along with a bounding box or segmentation mask that specifies the location and scale of the recognized object. For representation purposes, a visual vocabulary is used, i.e. an equivalent to a typical language vocabulary with an image corresponding to a part of a text. In the same way that text may be decomposed to a set of words, an image can also be decomposed to a set of visual words. Then, in order to compare two images, their corresponding visual words may be compared (Bag of Features or Bag of Words, BoW). The BoW representation aggregates visual information from independent and distinctive image patches, while discarding their spatial distribution. To extend the BoW model and add information from the spatial distribution of the local features, Grauman and Darrell proposed the Pyramid Match Kernel [6], which was used by Lazebnik et al. in Spatial Pyramid Matching [7] to create a series of BoW representation vectors corresponding to different regions of the image, in different levels of detail. The result retains both visual information and spatial distribution of distinctive image details.

In the direction of object detection, many algorithms, such as the Viola and Jones detector, exploit the sliding window approach. A virtual window of alternating size and aspect ratio slides throughout the image, examining a hypothesis of an object being depicted in every examined window. Using the sliding window scheme, Dalal and Triggs [8] proposed computing Histograms of Oriented Gradients (HOG) on the whole image instead of Haar features, describing the magnitude and angle of local image gradients. HOG features capture the boundaries of objects contained in images. Avoiding the sliding window approach, Felzenszwalb et al. [9] proposed a deformable part model, where an object consists of parts complying with a spatial distribution. In that case objects are detected by searching for local maxima of the responses of part filters.

Currently, the state-of-the-art methods for object recognition use Deep Neural Networks and in particular, Deep Convolutional Neural Networks (CNN) [10]. By efficiently exploiting big datasets such as Imagenet (1.2 million images) and high performance GPUs for fast computation of convolutions, learning deep CNNs - consisting of millions of parameters- in reasonable time became feasible.

The main part of our system multimedia repository consists of grayscale videos of standard and high definition. Most methods for generic object classification or detection using deep CNNs are trained using the Imagenet dataset, which consists of color images, where often color by itself is enough for distinguishing different object categories. Currently, there are no grayscale datasets of that size available, in order to exploit only the luminance of natural images for classification tasks. We therefore convert all images of Imagenet to grayscale and use them as training set for our deep CNN classifier.

To adopt the Deep CNN networks trained for image classification in object detection tasks, a localization step is needed. Using a sliding window approach leads to millions of windows being evaluated by the CNN classifier, which is computationally prohibited. In order to keep the computa-

tional cost low, a fast step of predicting windows that have a high possibility of depicting objects of interest is needed. Cheng et al. proposed BING, a very fast convolution-based method to select promising -in terms of the possibility to contain an object- subwindows, which proved to be effective [11]. Alexe et al. proposed objectness, a segmentation-based method for selecting subwindows, by checking the inclusion or intersection of image edges with the borders of candidate windows [12]. Uijlings et al. proposed Selective Search which relies on the same principle with Objectness, while combining different color spaces with different similarity measures for segmentation grouping [13]. In figure 4 we see different subwindows of an image that have high probability of containing objects of interest, using Selective Search.

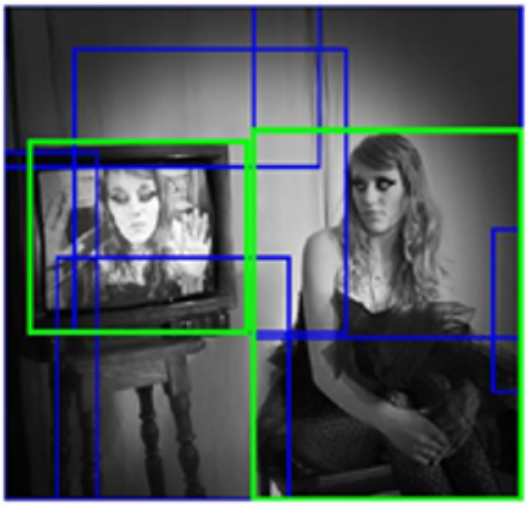


Figure 4. Proposing candidate subwindows, using Selective Search.

The generic object detection method we propose exploits the aforementioned technologies. We first use Selective Search for extracting candidate image subwindows for every image frame of the video in the repository. A pair of qualitative measures are applied on the results of Selective Search in order to keep only well-formed image regions. We have trained a deep convolutional neural network using the grayscale converted images of the ImageNet large scale dataset. The topology of our network is a replication of the AlexNet topology of Krizhevsky et al.. We adapt the input layer to handle grayscale images and discard color information. Then, we properly resize the proposed subwindows and feed them to our trained deep CNN. We keep only the results scoring above a given threshold value.

## 4. Landmark detection and recognition

The literature on local feature detection is rich and, since the early works based on the Hessian and the second moment matrices, many detectors have been proposed grounding on similar or novel ideas. The recent trend of achieving a good balance between efficiency and performance has led to a group of computationally efficient detectors. Indicatively, among them is the well-known SURF [14], an approximate version of SIFT, FAST [15], introducing fast corner detection based on an intensity comparison test in a small neighborhood and BRISK [16] that builds on the FAST detector.

Another group of methods, to which our group has contributed, exploit the stability of image edges by combining it with automatic scale selection [18]. For efficiency, these methods start from densely sampled edge points, compute an edge related function and detect regions by grouping its local maxima. There are also methods that exploit gradient strength directly, without edge detection. Avrithis et al. start from gradient strength and compute the weighted medial axis, decompose it into meaningful parts and group regions by taking both contrast and shape into account. Computational geometry concepts have been incorporated by our group [19] so as to propose more robust measures of stability and therefore increased performance. The Bag-of-Words model has been used as well. Avrithis et al. integrate appearance with global image geometry, while enjoying robustness against viewpoint change, photometric variations, occlusion and background clutter. All aforementioned enhancements, from query analysis to visual codebooks and geometry indexing, may be re-applied in all visual search related tasks [20].

In this work, we obtain automated location estimation and landmark recognition based on the state-of-the-art technologies described above. We implement a scene maps based methodology and have adapted it so as to enable condition-invariant (viewpoint, illumination, affine changes) landmark recognition for the available black-and-white content. Specifically, after preprocessing each frame and extracting local features we query the retrieval engine and retrieve similar photos along with their accompanying (when available) geo-tags. The core of the method is also available as a part of the standalone VIRaL[2] tool that offers richer interaction with the user.

## 5. Experimental study

The proposed system provides the above functionalities without user intervention, by automatically generating annotations in a per frame basis for videos. It exploits the technologies described above to extract information regarding the presence of objects, landmarks or people in images and videos. In our implementation, three frames per second are being extracted for processing, although a frame-by-frame approach can also be applied. Individual frames are visually enhanced by appropriate image processing techniques for noise reduction and contrast enhancement, before the pipeline reaches the analysis stage. All three analysis tasks are independent from each other and can run in parallel, using the same frame as input. For each analyzed frame, its timestamp is recorded. If a detected face is recognized as belonging to a person of the popular people in the database

2. http://viral.image.ntua.gr/

of our content providers, we also provide its instance name. The same stands for the objects, where the instance corresponds to the type of detected object, and for landmarks, indicating which landmark was identified. Especially for landmark recognition we also save the geographical location if available, as well as the corresponding text tags. We provide localization of detected concepts in terms of a rectangular, upright bounding box.

The proposed system was overall tested on content which was provided by the project partner Istituto Luce Cinecitta [3] and consists of approximately 3000 standard definition (320x240) black-and-white videos of news reels and documentaries, mainly from the 60's and 70's. 20 of these videos were also provided in high definition (1920x1080). Sample frames from the content can be seen in figure 5.

Figure 5. Sample frames of the LUCE video collection available in the MECANEX project.

All implementations of the aforementioned algorithms and techniques have been tested on challenging datasets against relevant or state-of-the-art techniques and have shown satisfactory or superior performance. For instance, in [5], the Local Binary Pattern Histograms framework is tested using the CSU Face Identification Evaluation System [21] was utilized to test the performance of the proposed algorithm. The system follows the procedure of the FERET test for semi-automatic face recognition algorithms [22] with slight modifications. LBPH-based face recognition proved to be superior against other relevant techniques, such as the Bayesian intra/extrapersonal (BIC) and the Elastic Bunch Graph Matching (EBGM) face recognition algorithms.

In [13], the authors test the Selective Search approach against relevant and state-of-the-art subwindow proposal techniques, such as the sliding window search of [9] and the objectness boxes of [12] and showed overall better recall and mean average best overlap (MABO) of the proposed subwindow against the manually annotated ones. The recognition technique used in our generic object detection pipeline was also tested in [10], on its performance on the datasets of the ILSVRC-2010 and ILSVRC-2012 challenges and achieved overall better top-1 and top-5 error rates. Finally, tests on our landmark detection methodology are documented in [20]. The European Cities 1M dataset, namely a subset of Barcelona landmarks, was used for evaluation. On this, the scene maps based pipeline proposed in [20] exceeds baseline (bag of words) and two query expansion approaches in mean average precision (mAP).

As the available content was equipped with manual annotations of persons present in a particular video file, we also conducted a set of experiments to validate the performance of the face recognition pipeline. A number of factors determined the final experimental protocol. Due to the fact that the people the our system is trained to detect are a relatively small subset of the total number of known persons present in the videos, it would not be fitting to perform a single experiment on the whole dataset. Furthermore, famous persons can be present in a video without showing their face at all or at least not in a way that a face detection based algorithm can perform well. Such occurrences can be present in the manual metadata, but cannot be expected to validate the systems performance accurately enough. It was decided then that, when a well-known person is detected by the system, its presence should be validated in the metadata, and not the other way around. With this in mind, three different experimental procedures were followed.

In the first one, a random sample of 500 videos was selected. The purpose of this was to test the performance of the face recognition task in an unconstrained environment, where there is no guarantee that any of the persons the tool is trained to detect is present. In other words, this experiment will showcase the low false positive rate achieved by the tool. In these set of randomly selected videos, the automatic annotation tool had a total of 52 false positive results, in the sense that it concluded that a person is present in the videos, whereas the manual metadata say otherwise. This can be translated to a 10.4% false-positive ratio. Individually, for specific persons, the tools accuracy in this experiment varies. For instance, Konrad Adenauer is falsely detected 6 times in the total of 500 videos. In contrast, Renato Rascel and Giulio Andreotti are only falsely detected 2 times in the whole dataset.

The second experiment involved a set of 20 hand-picked videos (selected by LUCE), which purposely contained persons that the tool is trained to detect. This experiment showcases theability of the tool to detect the persons it is supposed to, if they actually exist. The same protocol was followed. If a person is detected, the detection is validated with the manual metadata. In this experiment, the tool shows an overall accuracy of 45%. Notably, specific persons yield better accuracy. For instance, Giuseppe Togni yielded 100% accuracy, as the metadata validated the detection all 3 times his presence was recognized by the tool. Other persons, such as Silvana Mangano, yielded accuracy as high as 50%, because her presence was recognized in 2 videos and, according to the metadata, one was false.

With the third experiment, the ability of the tool to detect specific persons in a higher quality dataset was put to the test. In this set up, an attempt to guide the annotation procedure by looking only for the persons that are actually present in the videos was made. However, the protocol of validating the tools results with manual annotations and not the other way around was utilized here too, in order to have the same metric as with the previous experiments.

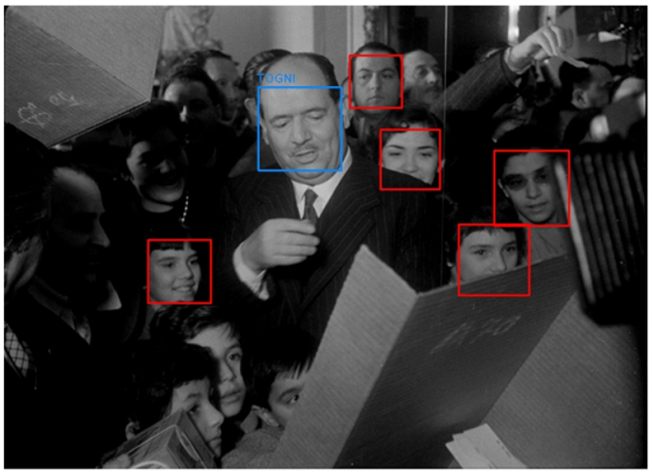
The dataset consisted of 10 high definition videos of the LUCE collection. The first five contained the following people: Giovanni Gronchi, Giuseppe Pella and Giuseppe Togni.

3. http://www.cinecitta.com/

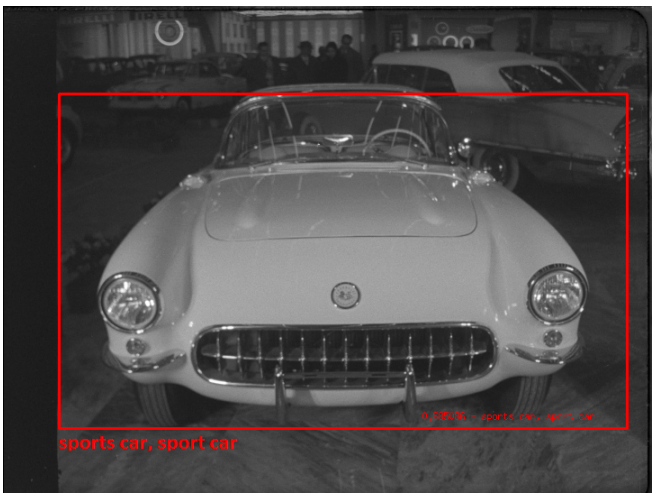
Gronchi was detected twice, when actually present in one video, yielding accuracy of 50%. Pella was not detected at all. Finally, Togni was detected in all videos he was present. This gives an average accuracy of 75% (detection/actual presence). The other group of five videos contained more people, namely: Giovanni Gronchi, Paolo Emilio Taviani, Emilio Colombo, Renato Rascel, Konrad Adenauer, Silvana Mangano and the Pope Pius XII. The tool achieved higher accuracy on the detected faces here, as the detections of three people (Gronchi, Adenauer and Rascel) were accurate every time. Taviani and Mangano were detected twice, while present in one video each. Finally, Colombo was detected three times, while only present in one video, yielding accuracy of approximately 33.34%. Subsequently, the average accuracy for that part of the dataset was 72.34%.

Given the fact that the majority of the occurrences are under bad conditions (object in the background, face not visible, etc.), the systems performance and especially the false-positive rate are overall deemed satisfactory. Another hindrance was the low quality of the majority of input data, which can be considered a natural factor and overall something that was under consideration from the very beginning. Some results can be seen in figure 6.

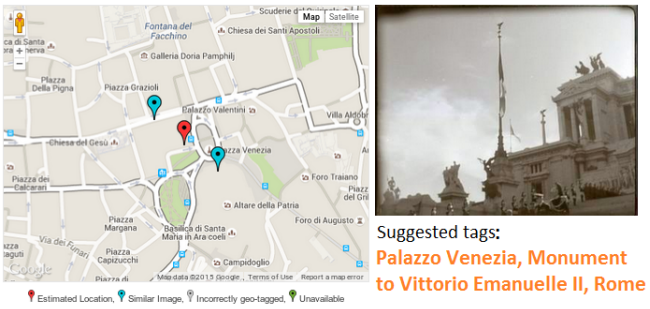
Image processing and analysis algorithms are in many cases computationally expensive. Complexity was a major concern when selecting the algorithms performing all image analysis tasks in the our automatic annotation Tool. The Viola and Jones method, as well as the Local Binary Pattern Histograms based face recognition run in real time in a modern CPU, having a low memory footprint. The complexity of the face detection process depends on the size of the frame, while the face recognition process is linear in the number of detected faces per frame. The generic object detection algorithm uses Deep CNNs, which are computationally expensive. To boost the time performance, the deep CNN is implemented using GPU accelerated code. This creates a hardware restriction for the machine running the object detection algorithm, as a high-end NVIDIA GPU is needed in order to maintain the high throughput. The complexity of the task is linear in the number of subwindow proposals given by Selective Search. Finally, for the landmark detection task, to keep the response time low for every query image, the representation histograms (in form of an inverted file), for all images in the database, have to be kept in memory, requiring approximately 8GB for our current database, consisting of 3 million images. The multimedia analysis algorithms used in the tool perform well in the application videos, providing high quality annotations.




(a) Detection of the Italian politician Giuseppe Togni.




(b) Detection of a sports car.




(c) Detection of a landmark in Rome.

Figure 6. Automatic annotation results

## 6. Conclusion

In this paper, an innovative pipeline of robust methods is presented for accurate automatic annotation of multimedia content in audiovisual industry scenarios. All techniques were implemented and configured to adapt to content of varying quality. Experiments and results show that this pipeline is suitable for generating annotations from audiovisual production videos and other multimedia. The system's integration in a real life framework and the utilization of its results in a number of use case scenarios involving real companies and industrial partners prove the accuracy and usefulness of the generated annotations as well as the necessity of implementing and leveraging such technologies.

# References


[1] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, Proceedings of Computer Vision and Pattern Recognition Conference, pp. 511-518, Kauai, USA, 8-14/12/2001.

[2] R. Lienhart, J. Maydt, An extended set of haar-like features for rapid object detection, Image Processing. 2002. Proceedings. 2002 International Conference on. Vol. 1. IEEE, 2002.

[3] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on. IEEE, 1991.

[4] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, IEEE Transactions on pattern analysis and machine intelligence 19.7 (1997): 711-720.

[5] T. Ahonen, A. Hadid, M. Pietikinen, Face recognition with local binary patterns, Proceedings of European Conference on Computer Vision, pp. 469-481, Prague, Czech Republic, 11-14/5/2004.

[6] K. Grauman, T. Darrell, The pyramid match kernel: Discriminative classification with sets of image features, Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on 17 Oct. 2005:1458-1465.

[7] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on 2006: 2169-2178.

[8] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on 25 Jun. 2005: 886-893.

[9] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Deva, Object detection with discriminatively trained part-based models, Pattern Analysis and Machine Intelligence, IEEE Transactions on 32.9 (2010): 1627-1645.

[10] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems pp. 1097-1105, 2012.

[11] M.M. Cheng, Z. Zhang, W.Y. Li, P. Torr, BING: Binarized normed gradients for objectness estimation at 300fps. Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on 23 Jun. 2014: 3286-3293.

[12] B. Alexe, T. Deselaers, V. Ferrari, Measuring the objectness of image windows, Pattern Analysis and Machine Intelligence, IEEE Transactions on 34.11 (2012): 2189-2202.

[13] J.R.R. Uijlings, K.E.A. Van de Sande, T. Gevers, A.W.M. Smeulders, Selective search for object recognition, International journal of computer vision 104.2 (2013): 154-171.

[14] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), Computer vision and image understanding 110.3 (2008): 346-359.

[15] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, Computer VisionECCV 2006 (2006): 430-443.

[16] S. Leutenegger, M. Chli, R.Y. Siegwart, BRISK: Binary robust invariant scalable keypoints, Computer Vision (ICCV), 2011 IEEE International Conference on 6 Nov. 2011: 2548-2555.

[17] Y. Avrithis, K. Rapantzikos, The medial feature detector: Stable regions from image boundaries, Computer Vision (ICCV), 2011 IEEE International Conference on 6 Nov. 2011: 1724-1731.

[18] K. Rapantzikos, Y. Avrithis, S. Kollias, Detecting regions from single scale edges, Trends and Topics in Computer Vision, pp. 298-311, 2012.

[19] C. Varytimidis, K. Rapantzikos, Y. Avrithis, S. Kollias, -shapes for local feature detection, Pattern Recognition, vol. 50, pp. 56-73, 2/2016.

[20] Y. Kalantidis, G. Tolias, Y. Avrithis, M. Phinikettos, E. Spyrou, P. Mylonas, S. Kollias, Viral: Visual image retrieval and localization, Multimedia Tools and Applications vol. 51.2, pp. 555-592, 2011

[21] D.S. Bolme, J.R. Beveridge, M. Teixeira, B.A. Draper, The CSU face identification evaluation system: Its purpose, features and structure, Third International Conference on Computer Vision Systems. (2003) 304311

[22] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, The FERET evaluation methodology for face recognition algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (2000) 10901104