Integration of Annotation Technologies in the PHAROS Platform

A Pharos whitepaper

Alessandro Bozzon, Marco Brambilla, Piero Fraternali

Web Models S.r.l.

Como, Italy

alessandro.bozzon, marco.brambilla, piero.fraternali @ webratio.com

May 28, 2009
Abstract

PHAROS is a configurable and customizable platform for Search-Based Applications (SBA). Such platforms allow integration of different kinds of annotation and indexing technologies for any media type. We report on an experiment of integration of an annotation component within the PHAROS configurable multimedia search platform. We integrate a third-party image analysis technology for human face recognition, by extending it in order to cope with the requirements of video annotation solutions and wrapping it according to the platform’s integration specification. As a result, the annotator outcomes are indexed within the PHAROS platform and can be used for multi-modal and content-based search. We evaluate our experience in terms of development effort required, indexing performances, and perceived quality of the annotation results.
1 Introduction

Considering the continuously increasing amount of information available on the Web, search is becoming the only feasible interaction paradigm for information seeking. Search-Based Applications (SBA) address this issue by providing search features over collections of heterogeneous data. Despite the common need of information, different profiles of users require different kind of search interfaces and searching mechanisms. For instance, end-user oriented thematic search portals are dramatically different from enterprise intranet search applications and from specialized audio content search systems for the media industry. Therefore, SBA present highly variable requirements, depending on the type of managed content and of content processing. However, most SBA share a common architectural structure, based on a complex front-end for query and result presentation, and a complex back-end for content provisioning, annotation, and indexing.

In this context, a few proposals of configurable SBA platforms exist that combine the advantages of solid and shared architecture with the advantage of complete configurability and/or extensibility of both front-end and back-end. A notable example is the PHAROS (Platform for searching of Audiovisual Resources across Online Spaces) platform\(^1\), developed within an project financed by the European Union under the IST 6th Framework. The PHAROS platform is completely configurable, allowing the definition of multiple front-end interfaces for every installation, and specification of flexible back-end processes. Among them, the most critical process is the Content Provisioning, Annotation, and Indexing (CPAI), that enables the retrieval, annotation, and indexing of any kind of distributed audiovisual content.

This paper aims at reporting on an integration experiment of external annotation components within the CPAI process. In particular, to demonstrate the extensibility and the plugin-based architecture of the platform, we leverage an existing component for human face recognition in images provided by KeeSquare S.r.l.\(^2\): we adapt it in order to meet the needs of the new usage environment, we wrap it as a Web service according to the integration specification of the PHAROS platform, and we adopt it within a CPAI process. Thanks to this activity, the annotators results are automatically indexed within the PHAROS platform and can be used together with any other annotator, for multi-modal annotations and/or content-based search. The experiment is interesting according to two aspects: (i) the integrated technology is introducing a completely new kind of annotation within the PHAROS platform, thus enabling its combination with other existing feature extraction components; and (ii) the existing face recognition technology is taken from a completely different scenario (real-time identification of blacklisted faces in still images for security purposes), extended to deal with video sources (movies, TV shows, news, documentaries, and so on), and enriched for providing new annotation information, such as segmentation of the video according to entering/exiting instant of people, clustering of similar faces that are likely to represent the same person, generation of best view snapshots for every identified person, and so on.

We report the result of our integration efforts of the face recognition annotators in terms of: satisfied requirements of the integration, development effort required for the integration, indexing performances of the extended annotator, and perceived quality of the resulting annotator.

\(^1\)http://www.pharos-audiovisual-search.eu/
\(^2\)http://www.keesquare.com
The paper is organized as follows: Section 2 summarizes the main characteristic of the PHAROS platform and of the KeeSquare face recognition technology; Section 3 describes the main steps of the integration work; Section 4 discusses and evaluates the outcome of the integration; Section 5 compares to the related works; and finally, Section 6 concludes.
2 Background

PHAROS is an Integrated Project co-financed by the European Union which aims at developing an innovative platform for multi-modal and content-based search of audiovisual contents [3]. The distinctive aspect of the PHAROS platform is its being an open framework for developing audiovisual search solutions, rather than a monolithic information retrieval system.

Figure 2.1. Architecture of the PHAROS platform.

Figure 2.1 depicts a high level view of the PHAROS platform, highlighting its main architectural components and interaction processes. All the modules in the architecture are designed according to the SOA paradigm, thus enhancing transparent functional distribution. The core element of PHAROS consists of an XML search engine, which operates on the annotations extracted from the content processing workflow, expressed in the Audio Visual RSS (AVRSS) format, an XML vocabulary extending the MPEG7 media description format [8]. All the other functionalities of the architecture are conceived to be pluggable: the modules in charge of registering the content, the feature extraction and content annotation algorithms, the workflow manager of the indexing process, are all designed for being flexibly added to the core framework of the architecture.

A fundamental issue in the design of PHAROS is concerned with the definition of the CPAI flow, whose goal is to effectively orchestrate the execution of a set of annotators in order to automatically generate descriptive metadata for the provided contents. Every annotator must be exposed as a Web service, it must comply with the platform’s data formats, and it must interact with the platform media storage components.

2.1 Morpheus SDK for Face Recognition

Face recognition and identification technologies are being more and more adopted in the multimedia indexing field. Such technologies are fundamental in professional information management (e.g., in security management, media industry, advertisement, and so on), and also in consumer information usage, as demonstrated by recent adoption in many end-user tools like Apple iPhoto or Google Picasa. In this work we adopt and extend Morpheus SDK, a software library engineered by Kee Square S.r.l.. The library offers a set of signal processing algorithms for biometric identification and tracking, detection of hazardous events, and security applications in general. Morpheus SDK technologies are able to efficiently process still images and real-time analysis of video streams. Morpheus SDK mainly focuses on facial recognition, performed through three subsequent steps:

- **Face Detection**: finds information on the position and size of each face visible in the image.
- **Face Normalization**: marks points of morphological interest throughout the face area (eyes, mouth, and eyebrows) in order to properly rotate and scale the image. Thanks to a proprietary implementation of the Statistical Models of Appearance method, the system identifies correctly morphological points even in the presence of severe occlusions (dark sunglasses, scarves, and so on) and different orientations of faces.
- **Feature Extraction**: selects relevant features from the normalized face, producing semantic, high-level descriptions, like gender (male, female) and age (baby, young, adult, elder). The implemented algorithms maximize the robustness against environmental disturbance and noise, non-optimal pose, non-neutral facial expressions, and variable illumination conditions.
- **Face Verification**: computes similarity values between the compared faces, by feeding the feature vector template (produced during the Feature Extraction step) to a statistical analysis engine. Using the trained statistical models, the feature vector is reduced to a smaller one that optimally describes the person’s identity, and then template differences are calculated.

Morpheus features an average processing time of 2 seconds for each video second, with a face detection rate of 99% for frontal faces with poses within 40 degrees from the center; the detection module is extremely fast thanks to the proprietary boosting algorithm with weak classifiers. High-level features are detected with a precision in the order of 85% for the gender, 75% for age, and 75% for the ethnic group.

---

2. www.apple.com/it/ilife/iphoto/
3. picasa.google.com/
4. All the reported values have been calculated by analyzing, on a server with 1 core unit working at 2.4 GHz) 24 hours of high-quality video produced by a major Italian television network. Analysis has been performed by keeping the minimum recognizable face size at the 10% of the video frame size.
5. The reported values have been calculated on a test bed of 1000 high-quality images, with faces holding a 75 pixel eye-to-eye distance.
3 Integration

Since PHAROS is an open platform designed according to the SOA paradigm, the CPAI workfow requires an annotation technology to comply with the following integration rules: (1) the annotator must be exposed as Web services; (2) it must communicate by means of standard SOAP messages; (3) exchanged information must comply with the platform's data formats for information exchange (i.e., MPEG7); and (4) it must interface with the platform's storage components in order to retrieve videos and store generated artifacts.

In the next two sections we report the main integration steps needed in order to (1) adapt the Morpheus SDK to the analysis of videos and (2) integrate the resulting annotator in the PHAROS CPAI workfow.

3.0.1 Extension of Morpheus SDK

The usage of Morpheus SDK requires some updates and adaptations in order to be able to efficiently address the field of video annotation and indexing. Two main additional requirements must be addressed:

1. **Redundancy control**: in a typical video with scenes containing many persons for a long time (e.g., interviews), a great deal of redundant information is produced. For instance, the same person may appear in many consecutive frames, possibly with different expressions or in different locations. In addition, the same person may enter or exit scenes within a video. The annotation technology must be able to effectively and efficiently reduce the amount of produced information by clustering annotations belonging to the same person, possibly by using a similarity measure to handle uncertainty brought by different poses, face expressions, or environment factors;

2. **Information completeness**: an annotation technology must grant to correctly (and completely) represent the visual content of the video. Such a requirement can be in contrast with the one of redundancy control: for instance, temporal information can be lost due to clustering of annotations. Correct annotations must include both correctly clustered features (e.g., unique faces) and temporal dimensions (e.g., the shots of appearance associated with the faces).

An effective annotation technology must manage the trade-off between these requirements, while offering additional functionalities like (1) **video segmentation**, based on the shots of appearance of people, (2) **people identification**, where given a database of known people, the annotation technology should be able to identify them, and (3) **video content abstraction and presentation**, where new media artifacts are generated from calculated annotations for effectively communicate the essence of the video to a user (e.g., pictures where a given person appears in the video).

Finally, non-functional requirements like performances, annotation quality, adaptability, and configurability should be addressed.

All the above-mentioned requirements have been tackled by our extension to the Morpheus annotator by means of a three steps video analysis process. Each step has been designed with the aim of minimizing the information redundancy while maximizing its completeness and precision. In this exten-
sion experiment other metrics like the computational cost have been partially addressed, although no specific optimizations have been implemented. Figure 3.2 shows the Video Face Analysis Process:

1. **Face detection and early clustering**: the input video is analyzed with the original Morpheus technology in order to extract, for each video frame, a normalized morphological template for each recognized face. Each face is then compared with the ones appearing in the previous frame in the video, so to create clusters of contiguous, similar faces. In this step clustering is only performed considering template similarity and positional contiguity in the video within given thresholds. For each resulting cluster, one representative template is chosen (together with its high level features such as age and gender), by taking into account its quality with respect to environmental noise, face expression, orientation, etc. This step can be configured through a threshold on the minimum size of the faces to be processed. Its output consists of a coarse list of faces, along with their best templates and their segments of appearance in the video.

2. **Final clustering and high-level feature extraction**: given as input the coarse list of faces, the second extended step performs a final face clustering for the whole video, where annotation for faces appearing in multiple shots are aggregated by considering template similarity. Then, for each face, a set of N-best representative templates are chosen. High-level features of the aggregate are then calculated starting from the ones of the aggregated faces, as a linear combination of the face template quality, their frequency, and their evaluation confidence. This allows to mediate possible wrong annotation evaluations associated with the chosen representatives (e.g., a bigger face with a non-optimal lightening can provide more precise information for the evaluation of the age of a person than a smaller one). This processing step produces the list of all the unique identities for the analyzed video along with their high-level annotations and their N-best templates.

3. **Person identification and derived artifact generation**: the last processing step tries to associate each recognized identity with a known person. Association is performed by comparing the identity's template(s) with the ones stored in the identity database: if the similarity within the templates is above a given threshold, then a recognized identity is labelled with the person's name, and the calculated high-level annotations are substituted with the a priori-knowledge about the matching person. Finally, derived artifacts for the identified people are generated (e.g., a representative video frame).

Figure 3.1 depicts an example of outcome for the video analysis process, where the topmost picture represents the best template for a recognize face (in this case, Al Gore) and the produced XML chunk contains the associated high-level features and segmentations.

The proposed method does not claim for optimality nor for annotation quality, but it represents a good compromise given the lack of a-priori information about the analyzed video collections. Thanks to the above-mentioned extensions, a technology born for still images and streams can be used for video annotations.

### 3.1 Integration in the PHAROS CPAI

The analysis process outlined in the previous section have been wrapped as depicted in Figure 3.2: the annotator services is invoked by means of a SOAP request which carries the URI of the video to analyze and the configuration parameters for the annotator. The video retrieval module is then in charge of

---

1 Thresholding a face's size is important because it helps filtering smaller faces, that are less relevant and provide less precise annotations.
downloading the video and, if required, transcoding it into a suitable format. The annotations are therefore transformed by the MPEG7 generation module into MPEG7 format. Finally, the Derived artifacts upload module uploads the generated artifacts in the Pharos content repository and the description MPEG7 is returned to the annotation workflow. To support fast incremental re-indexing of videos, the Web service also features smart caching mechanisms that can increase performances by storing copies of the downloaded (or transcoded) videos and generated MPEG7 description.

From the PHAROS platform’s point of view, integration only required (i) the extension of the annotation workflow with a new node for the KeeSquare module and (ii) the configuration of the aggregator for the different annotation technologies’ outputs in order to add face annotations and to exploit them for video segmentation.
Figure 3.1. Example of result for the video face analysis process.
Figure 3.2. High-level architecture of the developed Face Annotation Service.
4 Discussion

This section discusses the results of our work in terms of (1) required development effort, (2) indexing performances of the extended annotator, and (3) perceived quality of the results. The tests have been evaluated on a collection of 175 short video chunks, built from 10 of public speech videos taken from the TedTalks archive, that have been selected because of the high number of frames representing faces (usually the speaker, but also some other characters). Videos have an average length of 20 minutes and have been transcoded in H.264, with a resolution of 640x480 and 320x240 pixels. Each video has been split in chunks of 1, 2, and 4 minutes, thus producing 350 tracks. For each length, we randomly selected half of the resulting tracks.

Development Effort. The extension of the original annotator approximately required 10 days of two junior developers, while the integration of the extended annotator in PHAROS needed about 3 days. Figure 4.1 is a screenshot of the Pharos search engine interface including the face-based search resulting from the integration.

Indexing Performance. One of the main concerns in extending an existing technology to a new field or with new features is avoiding worsening its overall performances. Such concern was especially true for the extensions presented here, since they imply two cycles of post-processing.

<table>
<thead>
<tr>
<th>Min. face size</th>
<th>1 minute</th>
<th>2 minutes</th>
<th>4 minutes</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>154</td>
<td>321</td>
<td>1098</td>
</tr>
<tr>
<td>20%</td>
<td>80</td>
<td>290</td>
<td>328</td>
</tr>
<tr>
<td>30%</td>
<td>61</td>
<td>130</td>
<td>266</td>
</tr>
</tbody>
</table>

Table 4.1. Performances (in sec.) of the face annotator, for different video lengths (min.) and face size threshold (in % of the image size).

Table 1 reports the registered execution times of the extended annotator on a Dual Core machine, with an AMD Opteron 3Ghz processor and 3GB of RAM. As expected, the time required to annotate a video increases according both to the video length and to the minimum size of faces to be processed. For small videos, the analysis time has the same magnitude of the original technology. In the worse case, though, the average processing time is up to 8 times greater than the baseline technology, thus indicating that the extensions require further performance optimizations. Notice, though, that results are influenced by the continuous presence of the same speaker, that dramatically increases the redundant data and, hence, the clustering overload. Finally, an increase of the minimum face size leads to better performances (less faces are processed).

Perceived Annotation Quality. For evaluation of the quality of the produced annotation we consid-
Figure 4.1. GUI of the PHAROS search engine, including: 1) faceted search panel for Age and Gender annotations; 2) a face snapshot.

<table>
<thead>
<tr>
<th>Feature</th>
<th>10%</th>
<th>20%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified Faces</td>
<td>90%</td>
<td>95%</td>
<td>95%</td>
</tr>
<tr>
<td>Age</td>
<td>54%</td>
<td>62%</td>
<td>65%</td>
</tr>
<tr>
<td>Gender</td>
<td>79%</td>
<td>80%</td>
<td>82%</td>
</tr>
</tbody>
</table>

Table 4.2. Precision in the evaluation of the high-level features. Numbers represents the percentage of correct annotation values.

Considered the ability of the extended annotator both to effectively recognize all the faces in the video and to produce correct high-level annotations. Such evaluation has been conducted empirically by a pool of experts. Result are reported in Table 2: w.r.t. the original annotator (Section 2.1), the extended annotator maintains good precision in both identifying faces and in detecting people’s gender, with a small decrease in performances when smaller faces are detected, while age annotations perform considerably worse than the baseline technology. In order to better understand the motivation of the latter behavior, we conducted a more detailed inspection of the results, noticing a low annotation’s confidence. The issue was therefore due to the baseline Morpheus technology, whose training was not optimal for the collection.
5 Related Work

Video annotation for content-based indexing is subject to many research efforts, from multiple perspectives: image and speech annotations [11], object detection [5], etc. Face detection and identification techniques [1] aim to identify the regions in an image which contain a face, regardless of its characteristics and environmental conditions [10]. The extension approach described in this paper can be easily applied to any face detection system. Recent examples of face analysis frameworks are [4] and [12].

Our method also provides natural segmentation of video, that can be classified as a clip similarity method based on keyframe comparison [7]. The problem of reducing non-sequential keyframe extraction has been addressed extensively in the past [6] but, to the best of our knowledge, few works previously address both problems of redundancy control and information completeness by using face similarity (e.g., [9] use face detection based on skin color for fragmentation). Our work, instead, tackles both problems using face-related information.
6 Conclusions

We presented an experience in the extension of an existing face annotator and its integration within an audiovisual search platform. The integration was the first performed for the PHAROS platform and allowed to achieve fundamental experience in the field and improvements in the platforms. Current and future work include optimization of the performances of the extended annotator and an additional training of the original annotator to increase the results quality.
7 Disclaimer

This work is a whitepaper proposed and endorsed by the Pharos Consortium. PHAROS
\textsuperscript{3} (Platform for searching of Audiovisual Resources across Online Spaces) is an EU Integrated Project (IST-2005-2.6.3) of the EC IST 6th Framework. Its peculiarity consists in being an open framework for developing audiovisual search solutions. Every functionality of the architecture is conceived to be pluggable, according to the SOA paradigm.

This document is provided "as is", with no warranties expressed or implied. The Pharos Project, the Pharos consortium, and its partners are not liable in any case about the contents of this document. The copyrights are held by the authors of this paper. No warranties nor rights are granted to readers. No copies can be sold or distributed in any form without written authorization of the authors. All rights are reserved.

\footnote{http://www.pharos-audiovisual-search.eu/}
Bibliography