The 13th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST (2012)
D. Arnold, J. Kaminski, F. Niccolucci, and A. Stork (Editors)

Structuring and Embedding Image Captions: the V.I.F. Multi-modal System

Cristina N. Vasconcelos[1], Asla M. Sá [2], Marcio I. Sá [2] †, and Paulo Cezar P. Carvalho[2]

[1] Instituto de Computação, UFF, Brazil
[2] Escola de Matemática Aplicada, EMAp-FGV, Brazil

Abstract

Within the context of historical photographic annotated collections, we observe the frequent occurrence of some subsets of important characters, usually described in captions. For many years, image captions were annotated using natural language texts intended to be read by humans. Today, the information retrieval of structured information is appealing and the migration of natural language captions to structured information is desirable in a variety of photographic collections.

In this paper, we describe the Very *Important Faces (V.I.F.) system, which is designed to graphically document the occurrence of distinguished characters within photographic collections and store this information in a structured format useful for retrieval purposes. The V.I.F. system implements face detection in the image data and detects proper names in previously inserted captions if any are present. The user matches names to faces throughout the software interface in order to produce a photo annotation that is stored considering structured information principles. Once the matching is done, an efficient verification tool is proposed, which helps the expert to review the annotation, taking advantage of such multi-modal databases. The concept of annotation maturity level is also introduced.*

Categories and Subject Descriptors (according to ACM CCS): H.2.8 [Database Applications]: Image databases —Data mining H.5.2 [User Interfaces]: Training, help, and documentation—I.5.4 [Applications]: Computer vision —Text processing I.3.4 [Computer Graphics]: Graphics Utilities—

1. Introduction

Over the last decade several photographic collections have been digitised and many of them have been made available for public access through web portals, for instance see the Library of Congress' photo-stream on Flickr [Lib]. Each image may potentially have captions and/or texts that have been produced by experts to describe its content, which is usually stored as free text within a data basis. Captions may refer to the picture as a whole and/or describe a specific important feature that occurs in a particular subregion of the image. In order to specify the referred subregion with natural language, sentences like: *on top of*, *on the right of*, *in the first plane*, *from right to left*, *dressed in white*, *using a red hat*, etc. are used frequently. All of these sentences suffer from lack of precision, even from ambiguity in some cases, and their automatic processing can be difficult. Today several picture managing software, as well as social networks sites, make tools available for graphically annotating subregions of an image and associate it with a tag. The discussion in this paper is about how to bring these resources to the archive community in a meaningful way. As a result of this discussion we present the Very Important Faces (V.I.F.) system, which is composed of two modules: the annotation and verification module. The annotation module has been previously described in [SC] as a project paper, the achieved results and further discussion that lead us to the verification module is the original contribution here.

The primary goal of the proposed multi-modal annotation module is to provide tools to graphically annotate a set of

† Supported by CNPq-Brazil

DOI: 10.2312/VAST/VAST12/025-032

character faces and match them to their proper names, retrieving from a free text caption, as illustrated in Figure 1. The image shown in the example is part of the case study collection from CPDOC/FGV archive [CPD], which consists of a catalogue of contemporary historical figures organised by the archive experts. The case study collection and its particularities will be described in Section 3.

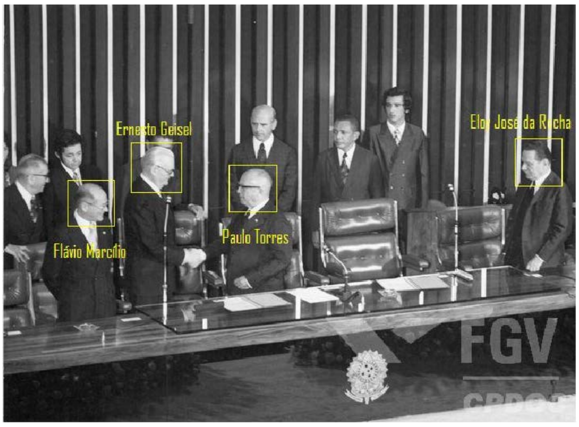


Figure 1: *Original caption: Esq./dir.: (1o plano/plane) Flávio Marcílio (1o); Ernesto Geisel (2o); Paulo Torres (3o); Eloy José da Rocha (4o). (2o plano/plane) Adalberto Pereira dos Santos (1o). Foto: Agência Nacional (Estúdio/Agência).*

In order to achieve the desired result, faces of important historical figures need to be detected within the images. Names should be detected from within the caption's text and the matching of names and faces has to be solved. The output information expected by the annotation module consists of descriptions of the faces' spatial positions within the images, followed by the name matched to the face. We call attention to the fact that the face's spatial position annotation is more precise than the natural language annotation and can potentially solve some of the ambiguities possibly present in captions; the drawback is that spatial descriptors are easily readable by machines, but not by people. Thus we generate the demand of a layer of processing between the information and the user. The proposed V.I.F. annotation module will be described in Section 4.

A tricky discussion arose when we proposed a semi automatic annotation approach to the curators of the case study archive: curators considered the automatic annotation, as well as the non-expert annotators, as potentially inserting errors in an information data-basis, which was originally produced by experts. For instance, even if our semi-automatic name to face matching were to be capable of 95% of correct assignments, that would mean to the curators that they were dealing with a solution that potentially inserts new content in the information basis with at least 5% of error, which is usually unacceptable.

In addition, we observe that when dealing with cultural heritage database annotation, it is crucial to define a maturity level related to the produced annotation. The *annotation maturity level* is to be used, for instance, to decide if a document is ready or cannot yet be published. From this discussion we conceived the V.I.F. annotation verification module. The annotation verification module implements the concept of annotation maturity level described in Section 5 and supports an authorised expert with tools for efficient verification/review of the annotation produced by non-experts, which includes both man and automates. The proposed verification module explores concepts from information visualisation, data query and ordering, and also user hierarchy over the annotated information from the photographic catalogues. It is detailed in Section 6.

The face-to-name matching annotation followed by the maturity level of the information can be stored in a file or potentially be formatted according to structured data standards. Image embedded metadata seems to be the natural tendency for dealing with questions related to association, storage and transmission of the image content description and we discuss this subject in Section 2.2 and present our approach in Section 7.

The overall structure of the paper is summarised here: The next section is a literature review. The case study collection is described in Section 3. The *V*ery Important Faces (V.I.F.) Annotation Module is presented in Section 4. Then, we present the concept of annotation maturity levels in Section 5 followed by the description of the *V*ery Important Faces (V.I.F.) Verification Module in Section 6. Our approach of embedding caption information into the image file in order to move from natural language to structured embedded captions is presented in Section 7. Finally, conclusions and future work are summarised in Section 8.

2. Related Work

This section briefly describes some image content annotation and verification tools followed by the description of some experiences on the adoption of image embedded metadata as a content annotation solution.

2.1. Image Annotation and Verification

In recent years, several tools for character annotations were made available. For instance, the social networking service Facebook [Fac] provides a face recognition tool that helps people identify pictures of their own friends, providing tag suggestions when requested by the user.

Google's Picasa photo organizer and editing software [Pic12] provide a face recognition tool that helps the user add name tags to faces in photos. The Picasa's face recognition tool identifies similar faces among the user's photos and groups them into an "Unnamed People" album. After that, the user can review the grouping results and manually add a name to similar faces automatically grouped. If a face appears in the "Unnamed People" album but the user does not want to name it, it is necessary to manually move that face image into the "Ignored People" group. Once the chosen faces are named, the annotation can be used to retrieve photos of specific characters by querying the names. Facebook's recognition tool uses the user's social relations to define the search space for the candidates' names. Picasa's does the same, but based on the user's previous annotations. In our case of study, we are interested in getting clues from previously inserted image captions.

We observe a growing interest in methods that exploit existent multi-modal data since they can potentially alleviate the complexity of the image recognition tasks and also the need for manual annotation, which is a costly and time-consuming process.

The approaches presented by Berg et al. in [TLBF04, TLBLMF04, BBE*07] adopt caption processing for retrieving clues about the identity of the people in a photo. Motivated by the huge photo collection already tagged with captions from journalistic databases, the authors introduced a dataset consisting of approximately half a million news pictures and captions collected from the Yahoo News website. After applying a frontal face detector to the dataset, each face image region is initially associated to a set of names, automatically extracted from the associated caption. Subsequently, a face clustering procedure is applied for the discrimination task.

The work proposed by Guillaumin et al. [GMVS08, GVS09, GMVS11] explores textual information as a weak supervision source to improve the learning of recognition models. The author introduces novel approaches that involve metric learning, nearest neighbour models and graph-based methods to learn from the visual and textual data, and similarity metrics on the identities of the individuals. They report achieving state-of-the-art results on several standard and challenging data sets, and conclude that learning using additional textual information improves the performance of visual recognition systems. Motivated by their results, our approach follows some of their concepts.

By experimentation, we concluded that, for our historical photographic catalogue, off-the-shelf software annotation tools have performed below the expectations. We will discuss our experience in Section 3. The most evident limitation of the majority of the available photo annotation tools is that they were not designed to process information available in captions or texts produced by experts that describe the content of previously organised photographic collections.

2.2. Embedded Metadata

Once the spatial position of a face is annotated and matched to a name, this information has to be stored in some way. In Picasa, the tagged names and their corresponding positions within the image are stored in a proprietary format that can be accessed in the `picasa.ini` archive but are restricted to being used within Picasa's interface due to license terms. The stored information consists of one line for each identified character that contains the corners of the face's bounding box followed by a number that identifies the character name in a locally managed name's databases. In the case of Facebook, the information is stored internally and is not made accessible.

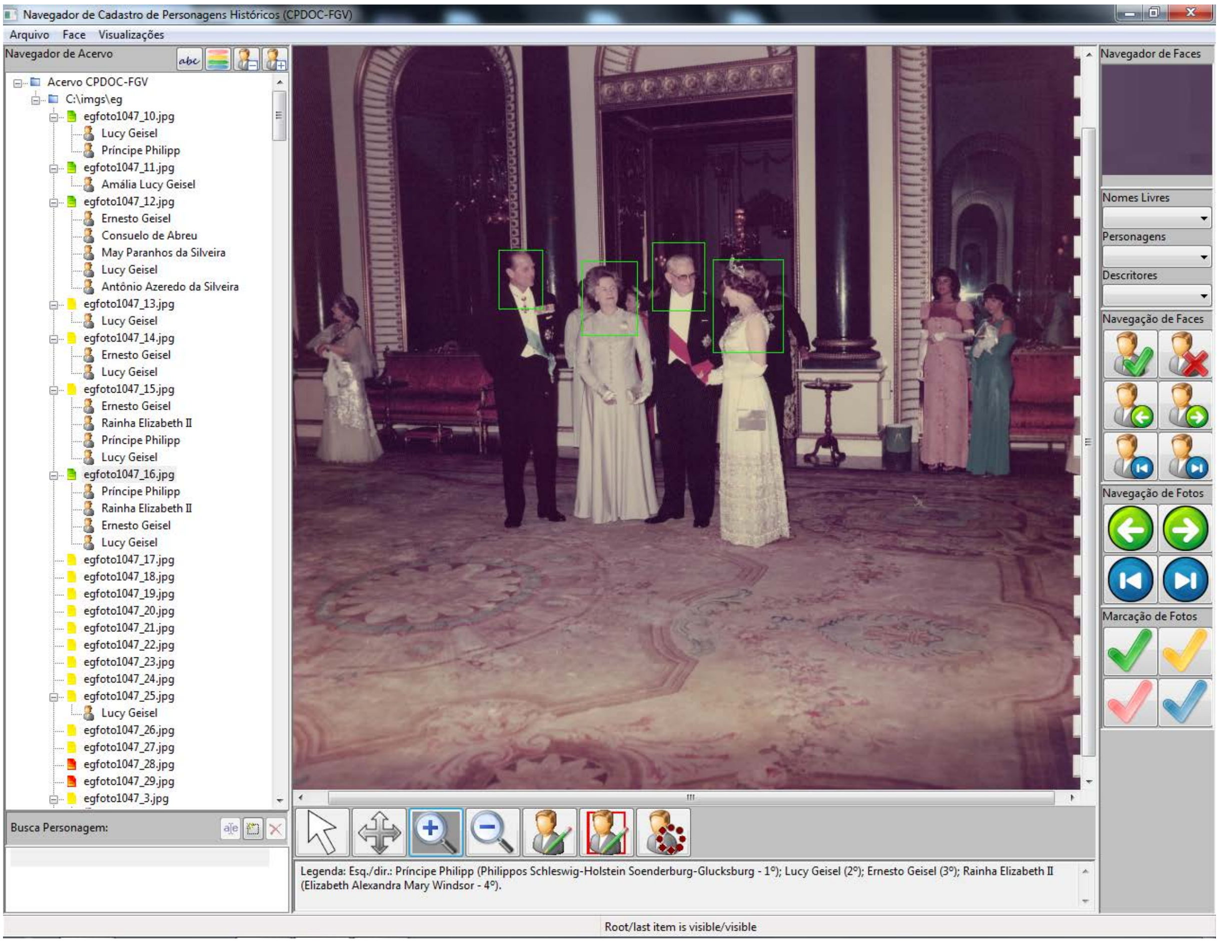
Our initial approach was to adopt a solution similar to Picasa's, but then we started to observe another tendency, which is to embed information in the image file by adopting photo metadata standards. The photo metadata standard is ruled by the IPTC Council [IPT] and is made to describe and manage photographs as well as to provide the most relevant rights related information. The standard is supported by major players in image industry like Adobe INC and telecommunication sites. Which primary goal is to make visual content accessible in human language terms or machine readable codes. In 2011 the IPTC Council launched the Embedded Metadata Manifesto [EMM] to argue how metadata should be embedded and preserved in digital media files.

The initial problem that IPTC Council intended to solve was related to image rights, but several examples of successful adoption of embedded metadata as a solution to store content description about the images are pushing the discussion of the role of photo metadata even further. Some examples of those are The Library of Congress' photostream on Flickr [Lib] and the Demotix news by you site [Dem]. Other examples can be found, but what we want to emphasise is that content regarding images can be embedded on the image file, and we are going to discuss this in Section 7.

3. Case Study: A Contemporary History Photographic Collection

The case study is a photographic collection consists of about 80,000 photos of contemporary history, from CPDOC/FGV [CPD] archive, that has been arranged and handled manually in its organization phase. In 2008, an extensive digitalisation project began, where the images and the results of the intellectual process of character identification and captioning were made available for public access through a web information portal. However, with the evolution of multimedia collection retrieval resources introduced by the use of semantic standards, the need to convey the collection to semantic standards arose.

The mentioned contemporary historical photographic collection has several important characteristics that were crucial to discard the use of off-the-shelf character annotation tools, since the adoption of such tools has been shown to be below the expectations of the experts in practice. Some of its particularities are the following:



- Acervo CPDOC-FGV
  - C:\imgs\eg
    - egfoto1047_10.jpg
      - Lucy Geisel
      - Príncipe Philipp
    - egfoto1047_11.jpg
      - Amália Lucy Geisel
    - egfoto1047_12.jpg
      - Ernesto Geisel
      - Consuelo de Abreu
      - May Paranhos da Silveira
      - Lucy Geisel
      - Antônio Azeredo da Silveira
    - egfoto1047_13.jpg
      - Lucy Geisel
    - egfoto1047_14.jpg
      - Ernesto Geisel
      - Lucy Geisel
    - egfoto1047_15.jpg
      - Ernesto Geisel
      - Rainha Elizabeth II
      - Príncipe Philipp
      - Lucy Geisel
    - egfoto1047_16.jpg
      - Príncipe Philipp
      - Rainha Elizabeth II
      - Ernesto Geisel
      - Lucy Geisel
    - egfoto1047_17.jpg
    - egfoto1047_18.jpg
    - egfoto1047_19.jpg
    - egfoto1047_20.jpg
    - egfoto1047_21.jpg
    - egfoto1047_22.jpg
    - egfoto1047_23.jpg
    - egfoto1047_24.jpg
    - egfoto1047_25.jpg
      - Lucy Geisel
    - egfoto1047_26.jpg
    - egfoto1047_27.jpg
    - egfoto1047_28.jpg
    - egfoto1047_29.jpg
    - egfoto1047_3.jpg

Legenda: Esq./dir.: Príncipe Philipp (Philippos Schleswig-Holstein Soenderburg-Glucksburg - 1º); Lucy Geisel (2º); Ernesto Geisel (3º); Rainha Elizabeth II (Elizabeth Alexandra Mary Windsor - 4º).

Root/last item is visible/visible

Figure 3: *The V.I.F. interface: the annotation and verification modulus are integrated with colours associated with both photos and faces tags, and also interface buttons. Therefore, users access and edit multi-modal content in a transparent manner.*

- the occurrence of non-frontal faces is very high (non-trivial for detection and recognition);
- typically the characters that are important within a single photo are just a few when compared to the number of faces that actually appear in the image;
- many of the images present some characteristic that makes the automatic image processing pipeline harder than usual; more specifically, may be monochromatic, contain different types of noise and may present the characters in very low resolution.

Figure 2 illustrates the common case where the detection of important characters when using Picasa's application fails (only faces from anonymous where actually detected).

4. The V.I.F. Annotation Module

The V.I.F. Annotation Module was designed to help archivists in the task of naming important characters that occur in photographic collections. Several state-of-the-art techniques were considered to automatically or semi automatically solve the problems faced. Some of these are: face detection, caption text processing and unsupervised approaches for matching faces and names in captions (Figure 3 presents a V.I.F. screenshot).

4.1. Face Detection

The *face detection* task consists of identifying subregions within an image where a human face occurs. Face detection is a well developed research subject and is already an off-the-shelf technique for the frontal face case. However, its general form is still a challenging computer vision problem due to variations in head pose orientation, facial expression, occlusions, variation in lighting and imaging conditions and the presence of non-uniform backgrounds.

Several face detection methods are available in the litera-

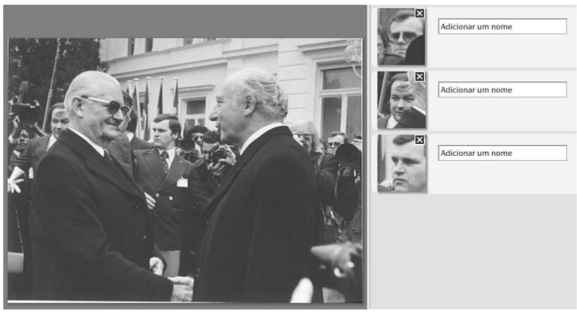


Figure 2: *Using Picasa on the case study collection: extras were detected, but characters that were considered important for the contemporary history database weren't, as they are not in near-frontal face poses.*

ture, see Jain's handbook [JL05]. Our system adopts the face detection framework that was initially proposed by Viola and Jones in [VJ04]. Which combines an efficient method for feature extraction (based on Haar Wavelets); a learning algorithm (Adaboost); and a cascade classification architecture. Their proposal can now be considered to be the "de facto" standard in face detection systems, having being adopted in many systems and digital cameras firmware and motivated by being computationally very efficient and fast. The V.I.F. Annotation Module adopts the Adaboost framework available in the OpenCV library [**?**]. However, motivated by the particularities of the case study collection, the classifier had to be retrained based on a training dataset prepared to deal with the historical image database variations of pose, expression and illumination. A great effort was put forth in defining and constructing the training data set properly.

It is important to note that although the face detection feature is available in the V.I.F tool, it can be turned off by the user to avoid cases where the presence of many unimportant characters (or extras) would make the work of deleting them more costly than the work of manually marking the important ones. In practice, if, for instance, Picasa were adopted, the work required to move extra faces classified as "Unnamed People" to the "Ignored People" album would be more costly than to name only the few people that are important in the case study collection.

4.2. Proper Names Detection

Concerning the textual information present in the captions, an automatic proper name extraction task had to be implemented in order to ease face tagging. In some cases, a natural language processing approach is demanded. For instance, in Berg et al. [TLBLMF04], a lexicon of proper names from all the captions was extracted, by identifying two or more capitalised words followed by a present tense verb. Words are classified as verbs by first applying a list of morphological rules to present tense singular forms, and then comparing these to a database of known verbs. This lexicon is matched to each caption.

While such lexicon construction was possible based on vocabulary and grammar public databases of the English language, in our case study a strong requirement of the archivists is that the created tags should belong to the controlled vocabulary established by a group of specialists. One of the reasons for such requirement is their precaution not to create ambiguous character names in captions, as there are common variants of names that refer to the same individual. Although this issue is usually solved by means of co-referencing, the controlled vocabulary available have had to be revised in order to incorporate co-referencing information.

Using such description dictionary, we extract proper names from caption, and this gives us a set of names associated with each picture.

4.3. Matching Faces to Names

In order to match the detected proper names to the detected faces our initial approach was to implement the graph matching approach proposed by Guillaumin et al. in [GMVS08]. The frequent presence of several extras in the collection images is the main reason why the graph matching approach does not work in the present challenging case study. Since the total number of names is usually much lower than the total number of automatically detected faces, a crucial assumption of the graph matching approach is violated. Therefore, we propose that the matching be manually done by the user, and we invested in the software interface to ease the user task.

5. Annotation Maturity Levels

In order to support the experts to efficiently review the annotation produced within V.I.F., which aims to avoid error inclusion in the database re-annotated by non-experts or by automatic processing, we proposed the adoption of a set of annotation maturity levels to tag annotation provenance. A maturity level tag is attributed to every annotation produced within the V.I.F. tool to indicate the annotation provenance and status. Maturity level tags are represented by different colours and can be associated to faces as well as to photos.

The maturity level tag attributed to a face represents the status in an face-to-name annotation process. The face maturity level tags in increasing order of maturity are respectively: red, that represents unnamed faces; yellow, that represents faces named by automates; green, that represents faces named by non-expert users; and blue, that represents named faces reviewed and approved by an expert. The tags attributed to faces are used to colour the face bounding box when the user is viewing a photo in the canvas. We call the attention to the fact that other tags can be added to the set

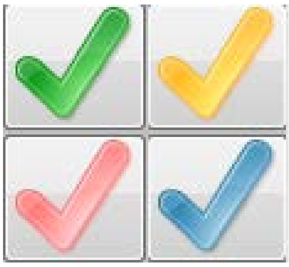
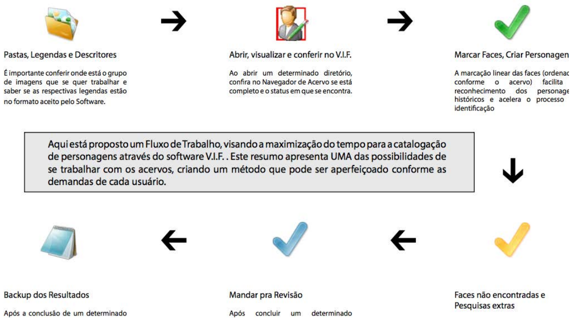


Figure 4: *Maturity level tags associated with a photo annotation status.*

of tags if other sources of annotation would be used, for instance, crowd-sourced annotation, or additional levels of review status could be incorporated if needed. By tagging the provenance of the annotation we are attributing a grade of confidence on the annotated information.

We also attribute maturity level tags to the photos, indicating its status within annotation work-flow (Figure 4). Each image maturity level tag is represented in the same way by colours, corresponding respectively, in increasing order of level of maturity: red represents un-annotated images; yellow represents on-going process on image annotation; green represents the images annotated by non-experts and ready to be reviewed; and blue represents images reviewed and approved by an expert, what means that the annotation is concluded.



Pastas, Legendas e Descritores

É importante conferir onde está o grupo de imagens que se quer trabalhar e saber se as respectivas legendas estão no formato aceito pelo Software.

Abrir, visualizar e conferir no V.I.F.

Ao abrir um determinado diretório, confira no Navegador de Acervo se está completo e o status em que se encontra.

Marcar Faces, Criar Personagen

A marcação linear das faces (ordenac conforme o acervo) facilita reconhecimento dos personage históricos e acelera o processo identificação

Aqui está proposto um Fluxo de Trabalho, visando a maximização do tempo para a catalogação de personagens através do software V.I.F. . Este resumo apresenta UMA das possibilidades de se trabalhar com os acervos, criando um método que pode ser aperfeiçoado conforme as demandas de cada usuário.

Backup dos Resultados

Após a conclusão de um determinado

Mandar pra Revisão

Após concluir um determinado

Faces não encontradas e Pesquisas extras

Figure 5: *Annotation-verification work-flow.*

Some of the maturity level tags are automatically attributed by the V.I.F. system during the proposed annotation-verification work-flow (Figure 5). When the user opens the directory containing the photo collection to be annotated, a red maturity level tag is automatically attributed to all the photos that have never been processed by our system. Once the user navigates through a photo, the face detection mechanism detects faces that are yet unnamed, thus the red face maturity level is attributed to them. Once this automatic processing is completed, the maturity level tag of the photo is changed automatically to the yellow tag, indicating that it is in an on-going status. The user can delete all the annotations at any moment by attributing the red tag to the photo, or can move forward by approving the face-names associations, by tagging the photo in green, meaning that the photo is ready for revision.

The blue maturity levels can only be attributed by an expert, and are protected within the V.I.F by a reviewer login. Once logged in, the expert can verify the annotations produced, approve in batch or change any photo or face maturity level Once a face level is downgraded, the photo level is automatically downgraded correspondingly. The concept of attributing colour tags to photos that corresponds to its review status is not a new one and is available in some photo management systems. The innovation here is the attribution of levels of confidence on the information based on its provenance.

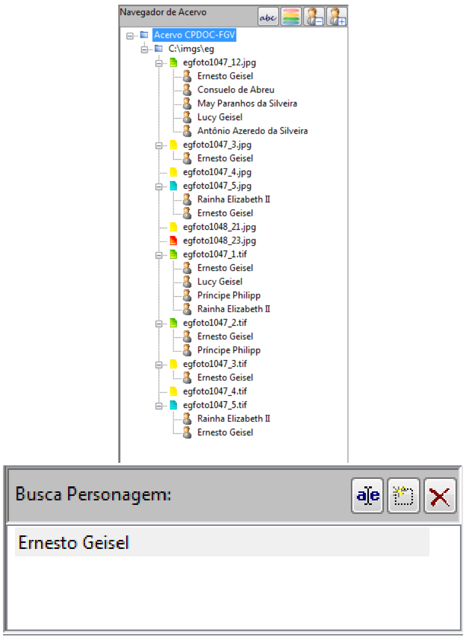
6. The V.I.F. Verification Module

The role of an expert is to guarantee a high level of confidence in the information associated with the photo, but experts can be expensive. Therefore, it is desirable to use this resource efficiently. Considering databases that have been previously annotated by experts, the migration of this information to structured standards would become infeasible if experts were required to redo the annotation task. Thus, less costly solutions should be proposed, that is, the task has to be performed by non-experts, crowd-sourcing or automates, leaving to the expert the task of verifying the annotation. In order to enhance efficiency in the verification work, some effort was applied in the V.I.F. design to offer querying and sorting tools over both the captions and the annotations. Interfaces based on fast information visualisation were also proposed so that an expert can quickly navigate through and review a summary of the annotation produced by non-experts. The following sections describes such tools in more details.

6.1. Querying and Sorting

The verification and the annotation of the photos can be done by querying a specific important character. For instance, in order to verify or to filter for annotation, all images within a collection containing a certain name in its caption can be searched for. In order to support the fast verification of a certain character, we have implemented a search tool that sweeps the captions from a photo collection selecting those photos containing the corresponding name. Once that is done, only those photos whose captions contain the query name are shown to the user. Figure 6 illustrates the proposed filtering by querying. The search tool can look for one or multiple characters.

Users can navigate through a photo collection, in any order desired. But, in some use cases, a set of photos is kept as

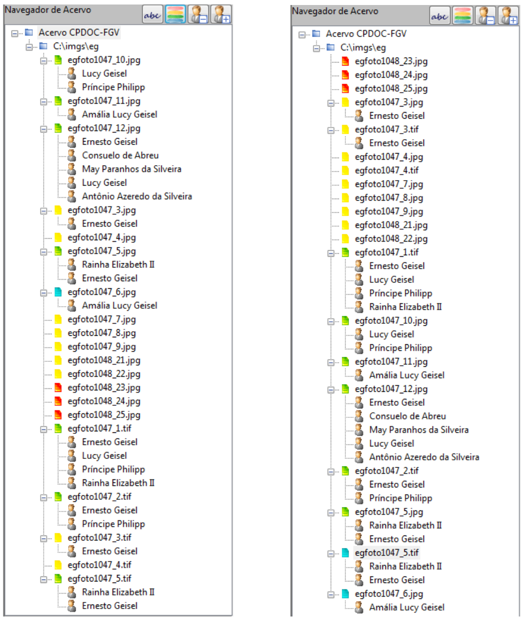


```
Navegador de Acervo
Acervo CPDOC-FGV
  C:\imgs\eg
    egfoto1047_12.jpg
      Ernesto Geisel
      Consuelo de Abreu
      May Paranhos da Silveira
      Lucy Geisel
      Antônio Azeredo da Silveira
    egfoto1047_3.jpg
      Ernesto Geisel
    egfoto1047_4.jpg
    egfoto1047_5.jpg
      Rainha Elizabeth II
      Ernesto Geisel
    egfoto1048_21.jpg
    egfoto1048_23.jpg
    egfoto1047_1.tif
      Ernesto Geisel
      Lucy Geisel
      Príncipe Philipp
      Rainha Elizabeth II
    egfoto1047_2.tif
      Ernesto Geisel
      Príncipe Philipp
    egfoto1047_3.tif
      Ernesto Geisel
    egfoto1047_4.tif
    egfoto1047_5.tif
      Rainha Elizabeth II
      Ernesto Geisel
```

Busca Personagem:

Ernesto Geisel

Figure 6: *Filtering photos by querying captions and annotations*



```
Navegador de Acervo
Acervo CPDOC-FGV
  C:\imgs\eg
    egfoto1047_10.jpg
      Lucy Geisel
      Príncipe Philipp
    egfoto1047_11.jpg
      Amália Lucy Geisel
    egfoto1047_12.jpg
      Ernesto Geisel
      Consuelo de Abreu
      May Paranhos da Silveira
      Lucy Geisel
      Antônio Azeredo da Silveira
    egfoto1047_3.jpg
      Ernesto Geisel
    egfoto1047_4.jpg
    egfoto1047_5.jpg
      Rainha Elizabeth II
      Ernesto Geisel
    egfoto1047_6.jpg
      Amália Lucy Geisel
    egfoto1047_7.jpg
    egfoto1047_8.jpg
    egfoto1047_9.jpg
    egfoto1048_21.jpg
    egfoto1048_22.jpg
    egfoto1048_23.jpg
    egfoto1048_24.jpg
    egfoto1048_25.jpg
    egfoto1047_1.tif
      Ernesto Geisel
      Lucy Geisel
      Príncipe Philipp
      Rainha Elizabeth II
    egfoto1047_2.tif
      Ernesto Geisel
      Príncipe Philipp
    egfoto1047_3.tif
      Ernesto Geisel
    egfoto1047_4.tif
    egfoto1047_5.tif
      Rainha Elizabeth II
      Ernesto Geisel
```

```
Navegador de Acervo
Acervo CPDOC-FGV
  C:\imgs\eg
    egfoto1048_23.jpg
    egfoto1048_24.jpg
    egfoto1048_25.jpg
    egfoto1047_3.jpg
      Ernesto Geisel
    egfoto1047_3.tif
      Ernesto Geisel
    egfoto1047_4.jpg
    egfoto1047_4.tif
    egfoto1047_7.jpg
    egfoto1047_8.jpg
    egfoto1047_9.jpg
    egfoto1048_21.jpg
    egfoto1048_22.jpg
    egfoto1047_1.tif
      Ernesto Geisel
      Lucy Geisel
      Príncipe Philipp
      Rainha Elizabeth II
    egfoto1047_10.jpg
      Lucy Geisel
      Príncipe Philipp
    egfoto1047_11.jpg
      Amália Lucy Geisel
    egfoto1047_12.jpg
      Ernesto Geisel
      Consuelo de Abreu
      May Paranhos da Silveira
      Lucy Geisel
      Antônio Azeredo da Silveira
    egfoto1047_2.tif
      Ernesto Geisel
      Príncipe Philipp
    egfoto1047_5.jpg
      Rainha Elizabeth II
      Ernesto Geisel
    egfoto1047_5.tif
      Rainha Elizabeth II
      Ernesto Geisel
    egfoto1047_6.jpg
      Amália Lucy Geisel
```

Figure 7: *a) Sorting by names. b) Sorting by maturity level status*

un-analysed to force further revision. Aiming to offer fast accesses to un-annotated photos (or any other maturity level), a photo sorting by name and also by maturity level tags was implemented Figure 7 illustrates both cases of sorting.

6.2. Visualisation

In order to easily verify every face image associated with a certain name, a visualisation tool is implemented so that the user can review an abstract of the annotations associated to a certain name. Through this visualisation the task of reviewing a collection annotation can be done character by character without the need to manually review every single photo.

None of the tools presented are new and they can be found in several other pieces of software, but vision of the reviewing task of the expert user as a quality controller of the information and the combination of tools to ease the control quality task is the contribution of the V.I.F. Verification module.

7. Embedded Structured Captions

Migrating captions in natural language produced by experts to structured text defined within a structured standards, to take advantage of automatic retrieval modules is a challenge. The volume of data in photographic collections is generally large. For that reason, there is great interest in unsupervised or semi supervised systems capable to migrate non-structured annotations to structured standards. State-of-the art algorithms for natural language processing and image processing can be quite useful for enhancing efficiency by automatically or semi-automatically executing a large part of the task. But still there is a decision to be made regarding how and where to store such structured information once it's been obtained. As we have already mentioned in 2.2, image embedded metadata is a tendency and is in the process of becoming strongly supported by the image industry due to demands related to the image's rights of use.

Since we have decided to adopt the IPTC embedded metadata standard [IPT], to annotate the V.I.F. output information, the first task to be done is to analyse the available fields of the standard in order to find the correct field to write each piece of information.

Sub-region annotation is a rich subject of discussion by itself. From a technical point of view, specification of sub-region can simply be done by defining two opposite corners of a rectangle in pixels, or any other preferred graphic description. Nevertheless, the corresponding structured information is not yet standardised. One possible reason for that is that annotation of sub-regions is a hard-to-read information for humans and, as we have already mentioned, a layer of processing has to be put on top of this information for it to become friendly to humans. This processing layer is the bridge between the precise hard-to-read sub-region annotation and the easy to identify visual annotation. Thus, the nature of the sub-region annotation data is slightly different from the data that is to be annotated in metadata fields that are usually supposed to be readable for humans.

In our proposal, we extend the fields of the standard to annotate the output information of the V.I.F. system, which consists of a set of names, their correspondent sub-region of occurrence within the image followed by the provenance of the information graded by its level of maturity tag.

8. Conclusions

In this paper we have presented the V.I.F. software developed as an environment for describing image contents throughout analysis, annotation and verification of multi-modal metadata.

The V.I.F. software combines face detection techniques and natural language basic processing to help the user in attribute names to faces that occur within an image through a user assisted interface. Currently, many applications offers the possibility to annotate contents within images subregion. For instance, several picture managing systems support face annotation by displaying a highlight to the corresponding subregion.

The V.I.F. software's singular contribution is that it has been designed to attend the requirements of a contemporary history multimedia database, that is, considering the specific demands of archivists that wish to annotate the occurrence of important characters in photographic collections. We argue that structured information is the approach to be taken to annotate the images. We incorporate into the V.I.F. tool the concept of user hierarchies relative to the annotation process, so that the provenance of the information and its level of confidence can be stored; The proposed work flow cycles between annotation and verification steps that considers both user hierarchies and annotation maturity levels to conclude the annotation flow. Finally, we argue that the adoption of the IPTC metadata standard is a feasible and interesting solution to the storage of the structured information produced within the V.I.F. system.

The first release version of the V.I.F tool is in the user test phase A large number of captioned images from the collection are being processed with the V.I.F. tool in order to produce the annotated information and embed it into the image file.

We believe that the proposed work-flow is useful in many other similar scenarios and that for each particular case specific image processing and automatic classifiers could be plugged into the V.I.F software in order to deal with the particularities of each collection.

9. Acknowledgements

This work is a result of great dedication of the archive experts that have made available the CPDOC/FGV - Arquivos Pessoais data basis for research purposes. We also thank CNPq/Brazil for supporting the participation of Marcio Isensee e Sá through the project EOS-Media.

References

[BBE*07] Berg T. L., Berg A. C., Edwards J., Maire M., White R., Teh Y. W., Learned-Miller E., Forsyth D. A.: *Names and Faces*. Tech. rep., U.C. Berkeley Technical Report, 2007. 3

[CPD] CPDOC/FGV: Cpdoc accessus - documentos de arquivos pessoais. http:http://cpdoc.fgv.br/acervo/arquivospessoais. 2, 3

[Dem] Demotix: Demotix: News by you. http:http://www.demotix.com/. 3

[EMM] EMM: Embedded metadata manifesto. http:http://www.embeddedmetadata.org/. 3

[Fac] Facebook: Facebook. http:http://www.facebook.com/. 2

[GMVS08] Guillaumin M., Mensink T., Verbeek J., Schmid C.: Automatic Face Naming with Caption-based Supervision. In *IEEE Conference on Computer Vision & Pattern Recognition (CPRV '08)* (Anchorage, United States, 2008), IEEE Computer society, pp. 1 – 8. 3, 5

[GMVS11] Guillaumin M., Mensink T., Verbeek J., Schmid C.: Face recognition from caption-based supervision. *International Journal of Computer Vision* (2011). 3

[GVS09] Guillaumin M., Verbeek J., Schmid C.: Is that you? Metric learning approaches for face identification. In *International Conference on Computer Vision* (Kyoto, Japan, Sept. 2009). 3

[IPT] IPTC: International press telecommunications council. http:http://www.iptc.org/. 3, 7

[JL05] Jain A. K., Li S. Z.: *Handbook of Face Recognition*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. 5

[Lib] Library of Congress: The library of congress' photostream on flickr. http:http://www.flickr.com/photos/library_of_congress/. 1, 3

[Pic12] Picasa: Picasa photo editing and web albuns, 2012. http:http://picasa.google.com/. 2

[SC] Sá A.; Vasconcelos C. G. M., Carvalho P. C. P.:. 1

[TLBF04] T. L. Berg A. C. Berg J. E., Forsyth D. A.: Whos in the picture. In *NIPS'04* (2004), pp. –1–1. 3

[TLBLMF04] T. L. Berg A. C. Berg J. E. M. M. R. W. Y. W. T., Learned-Miller E. G., Forsyth D. A.: Names and faces in the news. In *CVPR (2)'04* (2004), pp. 848–854. 3, 5

[VJ04] Viola P., Jones M. J.: Robust real-time face detection. *Int. J. Comput. Vision 57* (May 2004), 137–154. 5