The 12th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST (2011)
M. Dellepiane, F. Niccolucci, S. Pena Serna, H. Rushmeier, and L. Van Gool (Editors)

Very Important Faces: yet another character annotation tool

Asla Medeiros e Sá [1]

Cristina Nader Vasconcelos [3]

Martina Spohr Gonçalves[2]

Paulo Cezar Pinto Carvalho [1]

[1] Escola de Matemática Aplicada/FGV, Rio de Janeiro, Brasil
[2] CPDOC/FGV, Rio de Janeiro, Brasil
[3] Departamento de Computação - UFF, Niterói, Brasil

Abstract

This short paper describes the ongoing project of creating yet another character annotation tool, the Very Important Faces (V.I.F.) tool. Although the idea of character annotation is really not a new subject, off-the-shelf software annotation tools have proved to be designed for contexts where the assumptions are not the same as in the case of historic photographic catalogs. Thus, the adoption of such tools has shown, in practice, to be below the expectations. The most evident limitation of the majority of the available photo annotation tools is that they do not process information present in captions and texts produced by experts that describe the contents of the photographic collections. Our dataset is constituted of a contemporary historic character photographic collection, with informative captions, available for public access . The design proposal of the V.I.F. tool is to help the experts responsible for collection organization to migrate the information, documented in the texts associated to the images, to W3C metadata standards. The V.I.F Tool implements face detection algorithms. It also detects proper names in previously inserted captions to help the user (expert) match names and faces in order to produce a photo annotation compatible with semantic web principles.

Categories and Subject Descriptors (according to ACM CCS): I.3.4 [Computer Graphics]: Graphics Utilities—

1. Introduction

Over the last decade many photographic collections have been digitized and made available for public access through web portals. In many cases these collections have been organized and captions and texts have been produced by experts to describe the characters that appear in a given picture and the situation where it was taken. These portals are frequently referred as *I*nformation Portals [**?**] and act as hubs of data. The difficulty of finding a subset of images of interest in such portals is an obstacle for end-users.

The fact that the semantic web brings lots of solutions for the end-user to retrieve content is evident [**?**]. The challenge for the organizers of photographic collections is to migrate captions and texts produced by experts to the standards of the semantic web. The volume of data in photographic collections is generally large. Thus, there is a great interest in unsupervised systems capable to migrate information from non standard annotations to semantic standards.

The case study of the ongoing project described in this paper is a collection of about 80,000 digitized photographies of characters of contemporary history, organized by experts, that will be described in the next Section. Previous work will be briefly reviewed in Section 2. The *V*ery Important Faces (V.I.F.) Tool, designed to help the collection organizers experts to migrate information to W3C metadata standards will be presented in Section 4. Finally, conclusions and future work will be summarized in Section 5.

DOI: 10.2312/PE/VAST/VAST11S/017-020

2. Previous work

This section describes some recent applications that use automatic face detection and recognition techniques in order to support the task of querying tagged characters. Recent research on extracting semantic information from multimodal data associated to the image is also described.

In recent years several tools for character annotations were made available. The social networking service Facebook [FAC], as an example, provides a face recognition tool that helps people identify pictures of their own friends, providing tag suggestions when requested by the user. Other concurrent services also offer similar functionalities in their products.

The Google's Picasa photo organizer and editing software [PIC] provides 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 the 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" pile. Once the chosen faces are named, the annotation can be used to retrieve photos of specific characters more easily by querying the names. 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 it is restricted, by the software license terms, to be used only within Picasa's software interface.

The Facebookt's recognition tool uses the usert'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 tips from another kind of multimodal data. Since historical archives are typically labeled with captions, we look for technical approaches that explore such textual description in order to help character identification within the image.

There is a growing interest for methods that exploit existent multimodal data because 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 in [BBEF04, BBE*04, 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 created 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 Matthieu et al. [GMVS08, GVS09, GMVS11] explores textual information as a weak supervision source to improve the learning of recognition models. The author introduce novel approaches that involve metric learning, nearest neighbor models and graph-based methods to learn, from the visual and textual data, 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, we have adopted some of their solutions in our case of study.

3. The Contemporary History Photographic Collection

During the 38 years of the existence of the owner of the case study collection (the name of the owner was ommitted for this blind review version and will appear in the final camera ready paper) the photographic collection was arranged and handled manually. In 2008, an extensive digitalization 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 for migrating the collection to semantic standards arose.

The contemporary history photographic collection used as case study 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 shown to be below the expectations in practice. Some of its particularities are that the majority of the images are in black and white, tipically the characters that are important are few compared to lots of extra characters that appear in the image, and the occurence of non-frontal faces is very high. Figure 1 shows a case where the detection of important characters when using Picasa fails. We guess that the reason for this below expectation behavior is the fact that, nowadays, the majority of the common user's photos composition is centered on characters that matter with low occurence of extras and high occurence of frontal faces.

In practice, if Picasa were adopted, the work required to move extra faces clissified 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. The frequent presence of several extras is the main reason why the graph matching aproach proposed in [GMVS08] to associate detected faces to captioned proper names does not work in the present challenging case study.

Figure 1: *Using Picasa on the case study collection: lots of extras characters were detected instead of the characters that matter.*

4. V.I.F. Tool

In order to design the V.I.F. tool, that aims 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, fiducial point extraction, caption text processing and unsupervised approaches for matching faces and names in captions.

The *face detection* task consists of identifying subregions within a image where a human face occur. Face detection is a mature research subject and is already an on 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, character identity, occlusions, variation in lighting and imaging conditions and the presence of non uniform backgrounds.

Several face detection methods are available in the literature [JL05] based on several different technologies. An efficient face detection framework was initially proposed by Viola and Jones in [VJ04] as a combination of: an efficient method for feature extraction; a learning algorithm (Adaboost); and a cascade architecture. AdaBoost can be considered now as the "the facto" standard in face detection systems (adopted in many systems and digital cameras firmware) motivated by being computationally very efficient and fast. The efficiency of the framework is partially due to the great efficiency in discarding subregions where faces do not occur, that is, discarding non-faces subregions. The V.I.F. tool 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 consumed 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 lots of extras would make the work of deleting unimportant characters more costly than the work of manually marking the important ones.

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 general cases, a natural language processing approach is demanded. As an example, in [BBE*04], a lexicon of proper names from all the captions was extracted, by identifying two or more capitalized words followed by a present tense verb ([8]). 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 idiom, 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 exists commonly variants of names that refer to the same individual.

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

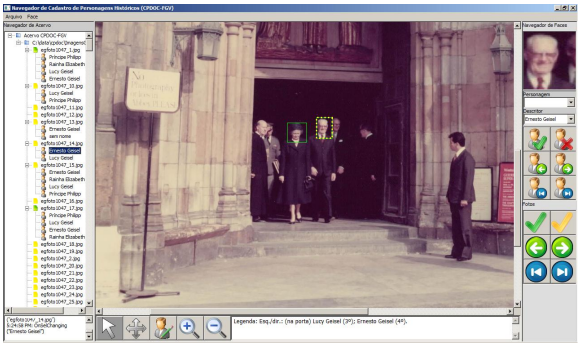

Figure 2: *The V.I.F. tool interface.*

The V.I.F. tool's interface is illustrated in Figure 2. The result of the tool processing is a set of character faces' bounding boxes associated to character proper names. This information is stored in a text file that can be easily formated according to W3C metadata standards, which is useful to produce nice visualizations of the information contained in the captions as illustrated in Figure 3. It also opens many new possibilities to retrieve information from the catalogs.

5. Conclusion and Future Work

In this ongoing project we designed and developped the V.I.F. tool, which is designed for the specific demands of

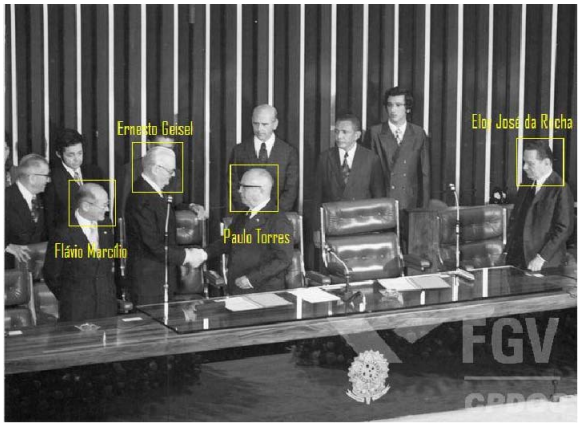

Figure 3: *Original caption: Esq./dir.: (1ž plano) Flávio Marcílio (1ž); Ernesto Geisel (2ž); Paulo Torres (3ž); Eloy José da Rocha (4ž). (2ž plano) Adalberto Pereira dos Santos (1ž). Foto: Agência Nacional (Estúdio/Agência).*

archivists that wish to annotate the occurence of characters in photographic collections acccording to W3C metadata standards.

The first version of the V.I.F tool is in phase of tests by the user. A large number of captioned images from the collection are being processed with the V.I.F. tool in order to produce the output information. The next step for the project in terms of research is to include in the tool a character recognitoin module intending to extend the photo annotation to new collections that were not captioned yet.

In order to achieve this goal, after the face detection phase is completed, the next technical step is to elect a description of those faces suitable to do discriminative comparisons between pair of faces, aiming at recognizing specific characters. Several methods look for robust structural features of a face, which should be constant among different conditions and variations of viewpoint, pose, or lighting. This is a promising scenario of research, and the richness of the available collection favor the possibilities to achieve good results.

References

[BBE*04] BERG T. L., BERG A. C., EDWARDS J., MAIRE M., WHITE R., TEH Y. W., LEARNED-MILLER E. G., FORSYTH D. A.: Names and faces in the news. In *CVPR (2)'04* (2004), pp. 848–854. 2, 3

[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. 2

[BBEF04] BERG T. L., BERG A. C., EDWARDS J., FORSYTH D. A.: Whos in the picture. In *NIPS'04* (2004), pp. –1–1. 2

[FAC] Facebook. Website. 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. 2

[GMVS11] GUILLAUMIN M., MENSINK T., VERBEEK J., SCHMID C.: Face recognition from caption-based supervision. *International Journal of Computer Vision* (2011). 2

[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). 2

[JL05] JAIN A. K., LI S. Z.: *Handbook of Face Recognition*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. 3

[PIC] Picasa photo editing and web albuns. Website. 2

[VJ04] VIOLA P., JONES M. J.: Robust real-time face detection. *Int. J. Comput. Vision 57* (May 2004), 137–154. 3