http://dx.doi.org/10.14236/ewic/sohuman2013.4

# Building social graphs from images through expert-based crowdsourcing

M. Dionisio, P. Fraternali,
D. Martinenghi, C. Pasini,
M. Tagliasacchi, S. Zagorac
Politecnico di Milano
Piazza Leonardo da Vinci, 32
20133 Milano, Italy
www.polimi.it/
lastname*@polimi.it*

E. Harloff, I. Micheel
J. Novak
European Institute for
Participatory Media
Wilhelmstr. 67
10117 Berlin, Germany
www.eipcm.org
*e.harloff@eipcm.org*
*i.micheel@eipcm.org*
*j.novak@eipcm.org*

**The extraction of semantic information from multimedia content represents a challenging problem. Despite the continuous refinement of automatic tools, the quality and completeness of the results is not always satisfactory. To overcome this limitation, the vision of the CUbRIK project is to provide a multimedia search and exploration platform that seamlessly integrate human tasks and algorithms. In this paper, as a concrete example, we illustrate the design of a multimedia content processing pipeline that aims at extracting evidence of social relationships from the analysis of a photo collection covering the main events and people that shaped the history of Europe after World War II. We discuss the issues faced by general-purpose crowdsourcing and automatic face detection/recognition algorithms in determining the identities of people portrayed in the photo collection. Hence, we illustrate the design of a system that tackles the uncertainty of the results produced by face detection/recognition with expert-based crowdsourcing.**

*Human Computation, Crowdsourcing, Digital Humanities, Face Detection, Face Recognition*

## 1. INTRODUCTION

Human computation is an approach to problem solving that integrates the computational power of machines with the perceptual, rational or social contribution of humans (Quinn and Bederson 2011). It can be applied to the resolution of distributed problems where neither the capability of machines nor that of humans alone suffice. One such problem is the detection of objects in images, for which machine algorithms still fail to provide a general-purpose solution with high accuracy.

Even the detection of faces and the recognition of face similarity, two of the tasks for which automated solutions grant good precision and recall, leave space to further improvement, because common algorithms typically work well under rather controlled conditions, most notably frontal face images and constraints on the minimum and maximum resolution.

In this paper we investigate an approach in which fully automated algorithms for face detection and identification are backed by human-executed tasks for boosting the accuracy of the automatic solutions by improving the input to the machine tasks. Unlike previous work in human computation for multimedia, e.g., object detection in image-tagging Games with a Purpose (von Ahn 2006), the proposed approach does not replace automated feature extraction by perceptual human work, but explores an architecture in which a *pipeline of tasks*, mixing machine and human processing, leads to the final results.

The application context of the proposed architecture is Digital Humanities (Schreibman et al. 2004), in which computer-based tools support humanities research. In particular, we aim at developing an application supporting the work of historians and librarians in cataloguing and putting into context visual historical materials, most notably photographs of historical events that contain a mix of identified and unidentified characters. The application should help the researcher to reconstruct the context in which the photo was taken, by identifying all the participants and the event represented by

the photo. For doing so, a mix of machine and human intelligence is adopted: *i)* photos in the input collection are scanned searching for persons (detected by their faces), *ii)* the faces corresponding with good probability to the same person are clustered, and *iii)* a historical social graph is built by exploiting the co-occurrence of people in photos. The co-occurrence graph acts as a tool to put an event in context, and gives hints to the researcher about the most probable event associated with the photo. At various stages, content in the collection can be enriched with content from the Web and the hypotheses (e.g., the identity of a person appearing in multiple photos) can be validated by the crowd.

Our main contribution are as follows:

• we present a general-purpose architecture for human computation problem-solving workflows, within the context of the CUbRIK Project (CUbRIK 2011);

• we instantiate such an architecture on the specific problem of content enrichment (i.e., annotation of multimedia content with semantic information) for digital humanities research, with an application for the *History of Europe* (HoE);

• we present the results of two different experiments: the evaluation of the utility of a non-specialized crowd for addressing face identification tasks; and the implementation of a completely automatic solution only based on machine components.

These experiments are preliminary to the next step, which is the engagement of a mixed crowd of experts and non-experts in the content enrichment pipeline.

## 2. THE CUBRIK PROJECT AND ARCHITECTURE

The CUbRIK project aims at developing a modular framework and distributed system architecture for flexible design and implementation of multimedia search applications allowing easy reuse of existing components and multimedia processing workflow, their extension with domain-specific elements, and the incorporation of human computation for tasks requiring human intelligence in the solution process. CUbRIK is a distributed system layered in four main tiers, as shown in Figure 1. The *Content and user acquisition tier* is responsible for registering content and users into the system. Users are of two kinds: searchers, who use CUbRIK applications for finding and interacting with information, and performers, who execute tasks (via gaming or Query&Answer) to provide contribution, semantic annotation, and conflict resolution.

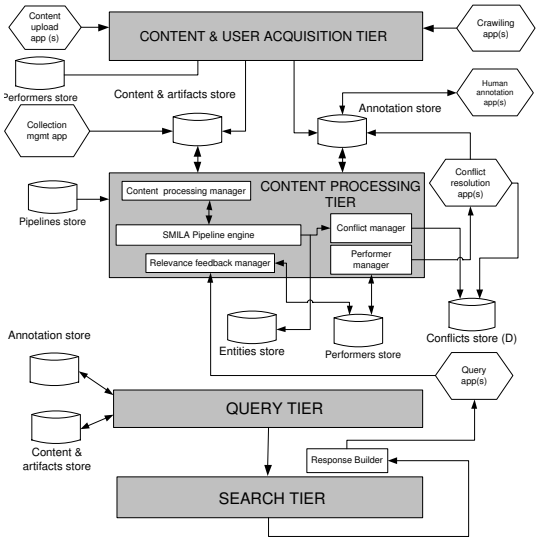


***Figure 1:*** *The architecture of the CUbRIK system*

In the *Content Processing Tier*, the *Content Processing Manager* listens to a queue of pending content processing requests and is responsible for managing tasks and processing contents at different levels of granularity. Process control is implemented on top of the SMILA pipeline engine (SMILA 2013) to orchestrate the execution both general-purpose (e.g., video/ audio segmentation) and domain-specific (e.g., face recognition) content-processing logic. The output of a content-processing task consists of: derivative content (e.g., key frames, thumbnails, audio summaries); low-level features, facts (i.e., annotations and confidence values); entities; and conflicts (i.e., low confidence facts and contradictory facts). The *Conflict Manager* is the core component for integrating human computation; it manages the set of conflicts and the assignment of conflicts to applications and performers. The *Performer Manager* is responsible for keeping data about performers, so as to optimize task allocation. Some pipelines are designed to receive feedback from the user on the results of a query. This feedback is routed to a *Relevance Feedback Manager* module that updates the level of trust of performers (human and automatic) in the component and performer store.

The *Query Processing Tier* consists of one or more *Query Apps*, which contain the front-end for issuing queries and viewing results; queries are expressed in a multimodal query language, serialized and submitted to a CUbRIK platform.

Last, the *Search Tier* contains a set of independent search engines that can access the content and annotation store(s).

## 3. USING GENERAL-PURPOSE CROWDS

In a first Proof of Concept, we designed and implemented a prototype for a face recognition service, which incorporated automatic mechanisms for face detection and recognition. The application context of HoE is challenging for purely automatic and algorithmic systems. In order to overcome this issue, a verification process for the automatic face recognition results by a general-purpose crowd via a crowdsourcing platform was introduced, in order to improve the underlying automatic results. We wanted to find whether the application of crowdsourcing for face recognition is able to improve or replace automatic solutions, especially in the case of heterogeneous data collections. Therefore, we wanted to evaluate the correctness of mappings of person names with snippets from group images out of the data set, which displayed only one person's face out of multiple faces on the full image. The data set also included portrait images which served as the basis for the creation of ground truth and the creation of human intelligence tasks for crowdsourcing. The image collection we used was provided by CVCE (Centre Virtuel de la Connaissance sur l'Europe) and was related to European heritage. Overall, the data set contained 3,924 images, 3,000 of which were used. For the collection, unstructured XML meta data was available, which enabled the creation of ground truth that was vital for the evaluation. The *expert-based ground truth* was created after a first iteration of automatic face detection, cutting and face recognition. A task set was set up which was solved by a historian from CVCE. He verified 574 associations of person names to snippets from group images, which led to a limited but necessary sample data set. For face detection and recognition, the 574 group image snippets were analyzed with methods from the (formerly free) face.com service provider. Since multiple candidate results (up to ten names in the case of face.com) are returned, crowdsourcing is introduced for effective result filtering, so that the size of the result sets can be limited to one or a few highly confident results. The execution of the human intelligence tasks (see example in Figure 2) was supported by the Microtask crowdsourcing platform. The workers affiliated with Microtask (Microtask 2013) are a general-purpose crowd, from Pakistan for this Proof of Concept, without expert domain knowledge. Each task was solved by ten distinct workers. A purely monetary incentive system was utilized, whereby workers earned approx. USD 4.00 per hour.

The tasks were designed as single image comparisons using two images and six possible user choices. The left image always displayed an already verified and annotated person on a portrait image.


***Figure 2:*** *Example of task for the History of Europe.*

The right image always was an automatically cut snippet from a group image displaying a person's face. For each snippet, up to ten tasks (one per candidate name) with different portrait images for comparison were created. In each task, a worker had the choice to either make undoubting or doubting statements about the comparisons. The workers' statements were then aggregated and evaluated using a proprietary majority voting schema (Harloff 2012). A probability mass function is constructed in three steps that narrows all the workers' answers to a certain probability that an image comparison task displayed the same person on both images. In advance, each worker's answer is mapped from a numeric rating scale $\{x \in \mathbb{N} | -2 \leq x \leq +2\}$ onto probability values in $\{x \in \mathbb{R} | 0 \leq x \leq 1\}$ . In the first aggregation step, a mean of each of the ten worker's answers is calculated. Secondly, if multiple portraits referring to the same person were used, the maximum mean of the step before is chosen. Thirdly, if the probability value is 0.5 or above, the related answer is considered relevant.

### Results of the first Proof of Concept

The prototype's solution approach using automatic face detection and recognition, plus crowdsourcing for result verification entails interesting results as well as strong limitations.

1. *Face verification results*. Of the 574 faces, only 17.1% were identified by the crowd, 66.0% of which were correct according to the ground truth, whereas the fully automatic baseline solution identified 80.4% correctly. Thus, in the specific application domain a general-purpose crowd tends to fail more often than algorithmic solutions. However, the crowd-verified results proved to be superior to the unverified automatic results. Indeed, the crowdsourcing results are almost always unambiguous (one result vs. up to ten) and effectively filter out false candidates. In conclusion, crowdsourcing-based verification proves effective for the filtering and conformation of face detection and recognition results. The prototype was strongly based on good automatic face recognition results by face.com. Thus, in case of no or false

automatic results, the verification was inherently incapable of finding the right result. Similarly for missing or falsely annotated portrait images, which were the basis for the comparison tasks. These issues are tackled in the second Proof of Concept (Section 5).

2. *Influence of image taking times*. For true positives (similarity of two correctly identified faces) as well as false negatives (similarity of two faces not identified), the time differences between the images were low, averaging 3.45 and 5.25 years, respectively. For false positives (similarity of two faces identified but false) this difference averages 13.5 years. We conclude that *i)* the smaller the difference, the easier the recognition, and *ii)* the bigger the difference, the harder and more inaccurate the recognition by humans.

3. *Limited size of Ground Truth*. The set of successfully expert-annotated group image snippets is vital and therefore affects the results. We found that even the CVCE expert could not recognize all persons. Therefore specific expert knowledge is needed to cover those cases in which a general-purpose crowd or automatic solutions are not good enough.

4. *Image resolution constraints*. For the face.com methods, a maximum image resolution was prescribed. For group images with many persons, this is a strongly constraining factor for successful and highly confident recognition.

5. *Replicability and trustworthiness of results*. The results are based on a majority voting algorithm and are neither author-based nor reasoned. Thus, the results of HoE users are not as trustworthy as those of experts. Also, the results on identifying European faces may depend on the specific cultural background and ethnicity of the workers and therefore may not be replicable in different settings. In the last section, we discuss whether selected expert crowds qualify to produce better results.

Despite the described limitations, we find that crowdsourcing generally enables the support and verification of automatic face recognition in the historic context examined.

## 4. CONTENT PROCESSING PIPELINE

In this section we describe the baseline workflow implemented to support the use cases of HoE. It consists of a completely automated sequence of multimedia analysis steps that produces a social graph from a collection of historic images. The purpose of this description is to highlight the accuracy of a purely automated approach, understand its weaknesses, and reveal where the injection of human intelligence tasks into the workflow has the highest potential for improvement.

The first phase of the pipeline is represented by the analysis of the whole HoE dataset by a state-of-the art face detection/recognition automatic tool (Kee Square 2013). The component is basically divided in two parts: the first devoted to face detection and the second to face recognition. The process begins with the face detection phase: the detector component receives as input a collection of images that are processed one by one. Once a photo is processed, the detector provides as output a collection of bounding boxes (regions of the image in which a face is detected). For each bounding box the component provides:

• Bounding box coordinates: coordinates of the top left and bottom right point of the bounding box.

• Additional info on the face 3d pose (roll, pitch, yaw).

• Detection quality: a number in [0,1] representing the quality of the detection of a bounding box. This can be interpreted as either *i)* the probability that a region of the image contains a face; or, *ii)* if the region contains a face, the quality of the info extracted from that face. This interpretation is crucial for face recognition, since the higher the detection quality, the better the computation of face similarity.

At the end of the detection phase, the face recognition component must be called because it can operate only on the bounding boxes, which are currently managed by the detection component, which passes them to its recognition counterpart. The face recognition component does two main things: 1. It extracts the biometric template of a bounding box detected in the first phase. 2. It compares two biometric templates in order to determine the similarity between the two faces that the two templates represent. Differently from the bounding boxes, the biometric templates can be stored and passed to the component later in order to perform face recognition. Once the biometric templates are generated, the component receives as input a pair of these templates and gives as output a number in [0,1] representing the similarity score between the two templates. In both phases some problems may arise (see Fig. 3), in particular:

• False positives: a detected bounding box is not a face (Due to the particular conditions of the images only the 40% of the objects detected by the tool are faces indeed). This issue is mainly related to the noise in the image (a bunch of noisy pixels can be

detected as a face) or to the high amount of details in the image (plants, clothes patterns, etc.).

• False negatives: a face in a photo is not detected by the component (The tool is capable to detect approximately the 75% of the total faces in the dataset). There may be several reasons: bad lighting in the image, occlusion around the face region, position of the face (e.g., particularly crooked faces are not detected).

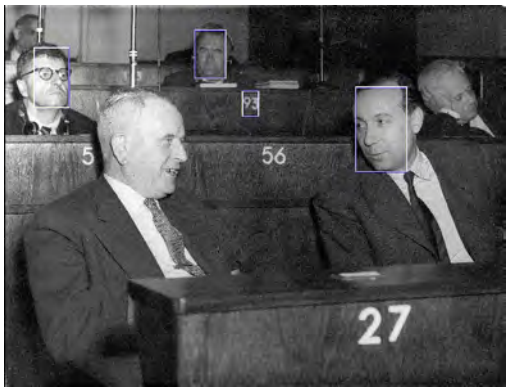


***Figure 3:*** *Typical output of the automatic face detector. Both the issues of false positives and false negatives can be seen.*

A possible solution to the issue of false positives is represented by analyzing the detection confidence provided by the tool: a threshold below which a bounding box is not considered a face and is thus discarded can be introduced, so as to reduce the occurrences of false positives. Reducing the false negatives requires preprocessing such as applying noise-cleaning filters and rotating the same image so that the undetected faces are put in a position useful for the detector.
In order to perform the identification phase, the dataset was enriched with portraits of characters present in the HoE photos and whose names were known. Those portraits were themselves processed by the detector and their bounding boxes were used for face matching with the ones detected in the HoE dataset. The main problem encountered in this phase is that the matching score between a character in a photo and his portrait is not always the highest one. This can be due to several factors, e.g., different person's ages in the two photos, different position of the face, different lighting, different face expression, etc. Some possible solutions to this issue can be: 1. Increasing the number of portraits of a single character (varying the face pose, the lighting, the age of the character and the face expression) can provide more faithful matching scores. 2. Crowd validations of the matches.

## 5. INTERFACING THE EXPERT-BASED CROWD

In addition to applying automatic face detection/recognition algorithms and involving general purpose crowds to provide information on the images from the collection, an expert-based crowdsourcing approach is explored. It attempts to overcome the deficiencies we encountered with general-purpose crowdworkers as discussed in Section 3. There, the crowd could only be used to verify results from automatic face detection and recognition as a filtering mechanism. While the results from the crowds were less ambiguous, their overall performance rate was worse. We believe using experts promises a much higher performance rate than the general-purpose crowd approach while also guaranteeing unambiguity and proper reasoning for answers, since experts can rely on their domain knowledge to identify persons and other contextual information (Heimerl et al. 2012). Experts are therefore also able to propose missing annotations rather than just verifying results from automatic preprocessing, thus exceeding the performance of automatic recognition. Consequently, designing tasks for experts should differ from general-purpose crowdsourcing task design.

The expert-based approach we propose is twofold, consisting of both explicit and implicit crowdsourcing. In the context of HoE, requirements analysis revealed that historians are already using existing social media networks, e.g. Twitter, to distribute mostly image-related queries such as "Who is this person?" among colleagues. This explicit expert crowdsourcing is based on community ties, as the participation incentives being at work, i.e., historians know they can rely on their colleagues to provide answers, which in return motivates them to answer. These existing expert communities within social networks should be utilized as the basis for a structured explicit query-based crowdsourcing solution. Therefore, we are developing an annotation tool for social media (see (Bozzon et al. 2012) for related work) that will enable historians to distribute inquiries via their established networks and easily retrieve and manage the answers their colleagues provide.

Secondly, we incorporate an implicit crowdsourcing approach similar to (Tungare et al. 2010). This requires image annotation and exploration tools that are seamlessly embedded in other research tools used by the experts assisting their daily workflow. A second Proof of Concept design visualizes such an annotation tool as it could be used for implicit expert crowdsourcing (see Fig. 4). With this tool, users can provide different suggestions for annotations, e.g., the name of a depicted person, or vote on existing suggestions for entities. The annotations and votes are also visible to other users, including explanations

for suggestions that were made and an indication of each user's level of expertise. Explanations, expert level indication of users and author-based majority voting allow for results that are much more trustworthy than those of the general-purpose crowd, which were only assessed using a probability-based majority voting algorithm. The suggested Proof of Concept design will be evaluated with users to get more insights on the most suitable task design and to verify the envisioned functionalities, e.g., the visualization of annotator agreement and mechanisms for collaborative conflict resolution.

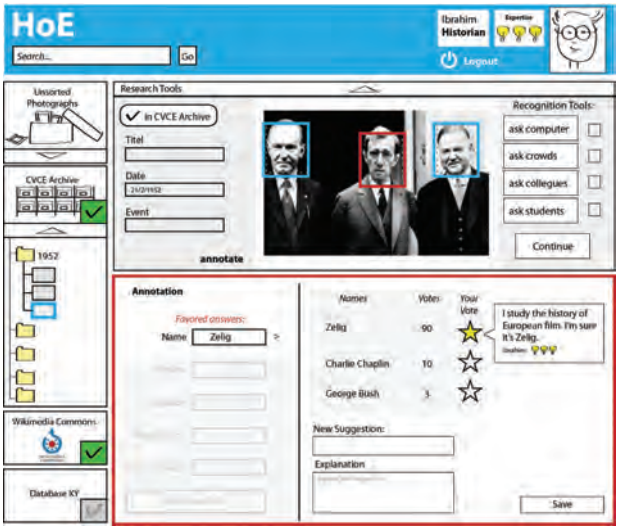


***Figure 4:*** *Implicit expert-based crowdsourcing Proof of Concept design.*

## 6. CONCLUSIONS

We have presented a human computation approach to extracting evidence of social relationships based on co-occurence of persons (face recognition) in photographs of historical events. The discussed findings from two experiments in face recognition in historical photo collections demonstrate the existence of specific challenges of the application domain (heterogeneity of content quality, differing scene compositions, the influence of the time dimension) that limit the applicability of both state-of-the-art machine algorithms as well as the applicability of general-purpose crowds. The proposed solution approach illustrates how the limitations of general purpose crowds (conceived as masses of unrelated individuals) in such domains may be addressed by designing an application that involves expert-crowds in different forms of collaboration in an interactive human computation system.

**Acknowledgments** This work is supported by CUbRIK (2011).

## REFERENCES

Bozzon, A., M. Brambilla, and S. Ceri (2012). Answering search queries with crowdsearcher. In *Proceedings of the 21st International Conference on World Wide Web, WWW 2012, Lyon, France, April 16-20, 2012*, New York, NY, USA, pp. 1009–1018. ACM.

CUbRIK (2011). EU FP7 CUbRIK Integrating Project, `http://www.cubrikproject.eu/` (retrieved June 25, 2013).

Harloff, E. (2012). Who is this person? Konzeption und prototypische Evaluierung einer Crowdsourcing-Anwendung für Multimedia-Suche. Bachelor thesis, Fachhochschule Stralsund, Fachbereich Wirtschaft.

Heimerl, K., B. Gawalt, K. Chen, T. Parikh, and B. Hartmann (2012). Communitysourcing: engaging local crowds to perform expert work via physical kiosks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, New York, NY, USA, pp. 1539–1548. ACM.

Kee Square (2013). Kee Square: Intelligent sensing for safety and security. Morpheus SDK, `http://www.keesquare.com` (retrieved June 25, 2013).

Microtask (2013). Microtask: Human powered document processing. Crowdsourcing Platform, `http://www.microtask.com` (retrieved June 25, 2013).

Quinn, A. J. and B. B. Bederson (2011). Human computation: a survey and taxonomy of a growing field. In *SIGCHI*, pp. 1403–1412.

Schreibman, S., R. Siemens, and J. Unsworth (2004). *A Companion to Digital Humanities*. Oxford.

SMILA (2013). SMILA: Unified Information Access Architecture. `http://www.eclipse.org/smila/` (retrieved June 25, 2013).

Tungare, M., B. Hanrahan, R. Quintana-Castillo, M. Stewart, and M. Pérez-Quiñones (2010). Collaborative human computation as a means of information management. In *Proceedings of the 2nd International Workshop on Collaborative Information Seeking at CSCW 2010*.

von Ahn, L. (2006). Games with a purpose. *IEEE Computer 39*(6), 92–94.