Expert Systems with Applications 42 (2015) 203–211

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Semi-automatic photograph tagging by combining context with content-based information

Hugo Feitosa de Figueirêdo[a,b,*], Cláudio de Souza Baptista[a], Marco Antonio Casanova[c], Tiago Eduardo da Silva[a], Anselmo Cardoso de Paiva[d]

[a] University of Campina Grande, Computer Science Department, Av. Aprígio Veloso, 882, Bodocongó – Campina Grande, Paraíba 58109-900, Brazil
[b] Federal Institute of Education, Science and Technology of Paraíba – Campus Monteiro, Monteiro, Paraíba, Brazil
[c] Pontifical Catholic University of Rio de Janeiro, Av. Marques de São Vicente, 225 – Rio de Janeiro, Rio de Janeiro 22451-900, Brazil
[d] Federal University of Maranhão, Applied Computing Group NCA, Av. dos Portugueses, SN – São Luís, Maranhão 58109-900, Brazil

ARTICLE INFO

Article history:
Available online 8 August 2014

Keywords:
Personal photo collections
Geo-referenced photos
People photo annotation
Metadata
Face recognition
Context-aware multimedia

ABSTRACT

This article proposes a semi-automatic technique for the annotation of people in photographs. The technique uses context and content information and is based on a weighted sum of estimators, which results in a list of the person's contacts that are more likely to be present in a photograph. Machine learning methods, such as multivariable linear regression and slope function, are adopted to filter and weight the estimators and eigenfaces for face recognition. The article also describes the results of experiments that were performed with a collection of 4050 photographs with 365 different people, which indicate that the proposed technique outperforms techniques that adopt only context or only content using as a performance metric the H-Hit rate of correct annotations.

1. Introduction

Approximately 350 million photos are added to Facebook each day, and during significant events, this number increases considerably; for example, on the weekend of January 1, 2011, over 750 million photos were added in a day. Whereas only a small portion of the captured photos are added to social networks, the number of digital photos reaches even higher figures. This large number of photographs involves an onerous task for the user in managing collections of digital photographs and, in addition, hampers the search through such collections.

Studies show that the annotation of information about the context in which a photograph was taken helps to retrieve it. In particular, identifying people who are present in the scene is considered to be the most important information that can enable a person to remember a photograph (Naaman, Yeh, Garcia-Molina, & Paepcke, 2005). However, most of the digital photographs are not annotated with such information.

The process of photograph annotation can be performed in three different ways: (1) manually, where the user performs the photo annotations himself; (2) semi-automatically, where an application helps the user to annotate the photograph; and (3) automatically, where an application annotates the photographs, perhaps with the help of external sources and content-based techniques (Escalante, Montes, & Sucar, 2012). Techniques to annotate people in photographs can in turn be divided into three groups: (1) techniques that use face recognition; (2) techniques that are based on the use of patterns of contextual information that are related to the environment in which the photograph was taken; and (3) hybrid techniques, which improve the performance of the face recognition algorithms by using patterns of contextual information.

Several semi-automatic techniques for annotating people in photographs have been proposed (Figueirêdo, Lacerda, Paiva, Casanova, & Baptista, 2012; Naaman et al., 2005; O'Hare & Smeaton, 2009). In particular, techniques that are based on contextual information typically use estimators to generate a recommendation of the people who are most likely to be present in a certain photograph. Those estimators are statistics that are based on contextual information of the previously captured photographs. Examples of estimators are: general recurrence of a contact in the collection of photographs; spatial recurrence of a contact in a certain area; and temporal recurrence of a contact in photographs that are taken within a time interval. Considering the hit rate of a

* Corresponding author at: University of Campina Grande, Computer Science Department, Av. Aprígio Veloso, 882, Bodocongó – Campina Grande, Paraíba 58109-900, Brazil.

E-mail addresses: hugo.figueiredo@ifpb.edu.br (H.F. de Figueirêdo), baptista@dsc.ufcg.edu.br (C. de Souza Baptista), casanova@inf.puc-rio.br (M.A. Casanova), paiva@deinf.ufma.br (A.C. de Paiva).

http://dx.doi.org/10.1016/j.eswa.2014.07.060
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

person who is present in both the photograph and in the recommendation list, the estimators vary according to their precision and recall. However, in most cases, combining several estimators leads to better overall results.

Automatic face recognition techniques, despite their low precision and recall (Pham, Maillot, Lim, & Chevallet, 2007), undoubtedly help to annotate people in photographs. Moreover, a semi-automatic annotation approach, which combines face recognition with contextual information, turns out to be a very attractive solution. Indeed, in applications that follow this approach, the user must manually inspect for annotation only those people who were not detected by the face recognition algorithm; furthermore, the application reduces the user's work by presenting a recommendation list from which people can be selected. In addition, the rate of correct recommendations can be gradually improved if the algorithm incorporates a feedback module that learns from the user's selections.

In this work, we define the content of an image as a set of image features that are generated from the pixel values. In the field of Content-Based Image Retrieval, CBIR, images are searched by their content rather than by using textual information (metadata). CBIR systems use statistics, pattern recognition, signal processing and computer vision techniques (Serrano-Talamantes, Avilés-Cruz, Villegas-Cortez, & Sossa-Azuela, 2013; Subrahmanyam, Maheshwari, & Balasubramanian, 2012; Yildizer, Balci, Hassan, & Alhajj, 2012). However, there is still a large gap between the content information that is used by CBIR systems and the semantics of the objects that are observed by the users in the images (Cambria & Hussain, 2012).

In some computer systems, information can be collected from the situation in which the user is placed to provide customised services and information. Such information is known as context, and it can be indexed for further retrieval. According to Dey (2001), context is any information that can be used to characterise the situation of an entity (e.g., photography).

In the domain of photography, the context in which a given photographic image was taken is very important for searching a collection of photographs. Examples of such context include the geographic location, date and time of capture, events, contact network and user preferences. For example, a user would like to retrieve the photos that were taken in July 2014, during the FIFA World Cup event, in which his girlfriend Ana was present.

In this article, we propose a semi-automatic technique for the annotation of people in photographs, which uses machine learning methods such as linear regression, slope function and eigenfaces. This technique combines contextual information and content analysis to generate a recommended list of people who are possibly present in a photograph. Through machine learning, the proposed technique weights and filters estimators, allowing new estimators to be dynamically added in such a way that the hit rate increases as the database increases. An estimator evaluates the probability that a person already identified by the system is present in a photograph. The technique uses two weighting methods, where the first uses linear regression and the second uses the slope function.

The main contribution of this article is that it proposes a semi-automatic strategy for the annotation of people in photographs and that the strategy has the following characteristics: (C1) it uses context plus content for annotation; (C2) it automatically computes estimator weights and proposes two weighting techniques to combine estimators; (C3) it incorporates an estimator filtering; and (C4) it uses linear regression, slope function and eigenfaces (methods based on machine learning) to improve the hit rate when new photos are annotated.

By context, we consider spatiotemporal metadata and list contacts, and by content, we mean face recognition. This article also describes the results of experiments that were performed with a collection of 4050 photographs, which indicate that the proposed technique outperforms techniques that adopt only context or only content, using as performance metrics the rate of correct annotations.

The remainder of this article is organised as follows. Section 2 describes related work. Section 3 overviews the proposed people photograph annotation method. Section 4 highlights an experimental evaluation of the proposed solution and discusses the results of the experiments that were performed. Finally, Section 5 presents the conclusions and suggestions for future research.

2. Related work

For the annotation of people present in photographs, the most common mechanisms use face recognition algorithms or contextual information. However, as was already noted, it is possible to combine both approaches. In this section, we cover some related work that uses content, context or both to perform the annotation of people in photographs.

Naaman et al. (2005) use the idea that there are patterns for the presence of people in photographs. These patterns are used to estimate the probability that a certain person is present in a photograph. The following estimators have been proposed: popularity, co-occurrence, and spatial and temporal recurrence. An application that follows this strategy suggests a list of people who have a high chance of being in a certain photograph by calculating these estimators. The probability of a person being present in a photograph is calculated as the sum of all of the estimators, using the same weight for each of them. In conclusion, this work uses a simple sum of the estimators; it does not account for the fact that some estimators could be more relevant than others. This limitation might reduce the accuracy.

PhotoMap (Viana et al., 2011) is a system that uses Bluetooth technology to annotate people who could be in a photo because they are near the photo capture place. Nonetheless, a portion of the annotated people could be in the photo, and others might not be. Thus, both the precision and recall could be affected due to erroneous people in the photo annotation.

MediAssist (O'Hare & Smeaton, 2009) is a system that offers browsing, search and semi-automatic annotation of personal photographs by analysing the content of the photograph and the context in which it was taken. The main semi-automatic annotation performed by MediAssist identifies people in the photographs. To build the list of people who have a higher chance of being in the photograph, MediAssist uses weighting techniques to combine the estimators. In the weighted combination, the weights were defined through a brute force approach, evaluating all of the possible values for the weights and adopting those that returned better results. Thus, MediAssist does not use machine learning to improve the results by utilising historical annotations. Furthermore, MediAssist does not filter estimators that will be used for each person who will be annotated.

MMM2 (Davis et al., 2006) is a system that uses a client–server architecture to perform the automatic annotation of people in photographs that are captured in mobile devices. MMM2 uses face recognition, data and time of capture, geographic location of the mobile device at the moment of capture, and people nearby. This last geographic context information is obtained from a mapping between users and Bluetooth. However, this system does not generate a list for semi-automatic annotation of persons in the photographs; it only uses the context to improve the performance of the face recognition algorithms. The temporal information is limited to Weekend or Weekday capture and hour of day. Additionally, the geographical information is limited to 100 groups created by clustering methods.