Multimed Tools Appl (2011) 55:7–26
DOI 10.1007/s11042-010-0586-z

Methods for automatic and assisted image annotation

Rui Jesus · Arnaldo J. Abrantes · Nuno Correia

Published online: 18 August 2010

Abstract Personal memories composed of digital pictures are very popular at the moment. To retrieve these media items annotation is required. During the last years, several approaches have been proposed in order to overcome the image annotation problem. This paper presents our proposals to address this problem. Automatic and semi-automatic learning methods for semantic concepts are presented. The automatic method is based on semantic concepts estimated using visual content, context metadata and audio information. The semi-automatic method is based on results provided by a computer game. The paper describes our proposals and presents their evaluations.

Keywords Personal memories · Media annotation and retrieval · Semantic image analysis · Tagging games

1 Introduction

Due to the advances in digital technology and the success of the World Wide Web as a platform for sharing media (e.g., using Flickr or Youtube), people are recording and sharing every moment of their lives. As a consequence, more experiences are preserved but it also becomes more difficult to later access this information to remember a relevant past moment. These difficulties arise mainly because personal pictures are being collected, most of the time, in a disorganized way without any type of annotation [9].

R. Jesus (✉) · A. J. Abrantes
Multimedia and Machine Learning Group, Instituto Superior de Engenharia de Lisboa,
Rua Conselheiro Emidio Navarro n 1, Lisboa, Portugal
e-mail: rjesus@deetc.isel.ipl.pt

N. Correia
CITI, Departamento de Informatica, Faculdade de Ciencias e Tecnologia, FCT,
Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Personal memories are recalled by humans through the episodic memory [34] which is related with the memory of events, times, places and other knowledge about a past experience. Due to the diversity of the clues provided by our memory, tools to search for personal media should provide ways to express these different types of information (query system) and the collection should be annotated [18] with these clues to support efficient retrieval.

The manual annotation of images with keywords describing their content is one way to efficiently provide media annotation. However, humans tend to avoid this operation [9, 36] because it is a time consuming task in large collections. The alternative is to develop automatic techniques to perform image annotation. The majority of the automatic methods are based on semantic models that are estimated using visual content [6]. This is one way to avoid the semantic gap problem [20] but as stated in the TRECVID 2006 overview [27] some difficulties remain.

Currently, most of the digital cameras have a built-in microphone. They save the camera parameters in the EXIF (Exchangeable Image File format) header of the JPEG (Joint Photographic Experts Group) file at capture time and some of them have integrated GPS (Global Positioning System) receivers to register the image location. It is expected that in the future more sensors will be included in the capture devices to record more information at capture time, in a similar way to the SenseCam. Therefore, image annotation methods should explore this contextual information in order to better understand the real world and consequently reduce the sensory gap [33].

This paper presents a semi-automatic image annotation method based on annotations provided automatically using semantic concepts and on user interventions to correct them through a computer game. Our strategy address the semantic gap with the feedback given by the users and with a semantic model that maps the low-level information in high-level concepts. The proposed method also address the sensory gap because images are analyzed by exploring their visual content and the contextual metadata (time and location) and audio information obtained at capture time.

The next section presents the related work and the following gives an overview of the system. Section 4, describes the multimodal image retrieval method. The semi-automatic application for image tagging is described in Section 5. The paper ends describing the tests conducted to evaluate our proposal and with the conclusions and directions for future work.

2 Related work

The image retrieval method proposed is based on semantic concepts that can be used to directly create a ranked list, given a query, or to annotate images with semantic meaning. We start this section by presenting and discussing several ways for image annotation. The term “automatic annotation” is used to describe the methods based on semantic concepts. The section ends with a description of related work about automatic and semi-automatic methods which are the categories where our proposals fit.

Different approaches have been proposed in order to annotate pictures with keywords describing their content. We classify them using the following categories: (1) manual annotation, (2) collaborative annotation, (3) annotation with recognized words using ASR (Automatic Speech Recognition) tools, (4) annotation using an entertainment application, (5) semi-automatic and (6) automatic annotation. Table 1 summarizes the main characteristics of these categories.

The manual annotation can be provided by choosing labels from a pre-defined set [31] or by typing words and associating them to an image (e.g., Picasa or Adobe Photoshop Album). As mentioned before, one of the weaknesses of this method is related to the human effort needed to annotate large collections of pictures [9, 36]. Manual annotation provided in a collaborative way (as it happens using Flickr) is more efficient. Although these annotations may contain noise due to errors generated by some users, several users annotating the same image contribute to a richer annotation set (requiring less human effort from each participant). Annotations obtained by recognizing words from audio files [30] recorded when the user speaks about their photos using a microphone (Table 1) require less user intervention. The problem is related to the recognition errors which can be frustrating to the user.

Previous methods demand some human effort but they are the most efficient. To provide fully automatic image annotation, systems need to extract information from the visual content or to use the camera metadata obtained at capture time and recorded in the EXIF header of the JPEG image file. This metadata provides useful information to retrieve pictures but to retrieve images with more complex information (e.g., people or buildings), the visual content must also be considered. The information used are the visual features automatically extracted or the semantic models estimated from these features [6, 20]. Nevertheless, the automatic image annotation is not as accurate as the manual process [27].

Semi-automatic methods attempt to solve the problem by including the user in the process [36]. They increase the annotation efficiency but they also increase the human effort when compared with the automatic annotation. The human effort needed by an application to perform one annotation plays an important role when the user is included in the task. In [37] time models are proposed for two manual annotation approaches to quantify the human effort.

Another option to annotate images is to turn the annotation process in an enjoyable task. In [35], this problem was addressed by replacing the manual annotation process with a computer game for online content. The human effort is the same but spent in a fun way. Since the annotation is manual the high performance is guaranteed.

Table 1 Main features for several annotation techniques: human effort, efficiency, input provided by the user and information used by the system

Annotation methods	Features			
	Human effort	Efficiency	Input	Information
Manual	high	high	text	keywords
Collaborative	medium/high	high	text	keywords
ASR Tools	medium	medium	audio	keywords
Semi-automatic	medium	medium	images	content or context features
Entertainment	low	high	text	keywords
Automatic	low	low	images	content or context features

2.1 Automatic methods

During the last years, several approaches have been proposed to automatically annotate images based on semantic models. Generative models [2, 21], machine translation models [7, 19], Bayesian networks [25], latent space models [23], hierarchical boosting algorithms [8] and based on an agent framework [3] have been some of the techniques used for automatic image annotation. An early approach for automatic image annotation was proposed in [24]. Mori et al. applied a co-occurrence model to words and low-level features extracted from rectangular image regions obtained using a regular grid. More recently, new approaches [4, 5, 16] in the domain of personal photos that integrates content and context metadata have been proposed. In [4], an approach is proposed based on probabilistic graphical models to recognize landmarks and people with location information and visual features. Temporal data combined with visual information is used in [5] to detect faces. In [16] it was proposed to use the metadata obtained at capture time (time, exposure time, subject distance and flash fired) with visual content for scene classification. This work has similarities with our proposal but they use SVMs for image classification and a different method based on Hidden Markov models to include the metadata. Additionally, GPS data and audio information is also included in our approach.

2.2 Semi-automatic methods

Several approaches [17, 22, 36] were proposed that use the relevance feedback mechanism for image annotation with keywords. Lu et al. created a semantic network with a set of words having links with weights to a set of images. These connections are based on the user feedback. They adapt Rocchio's formula to incorporate the low-level features and high-level semantic feedback. In [17], a system was proposed that is closer to our proposal but their system is based on SVMs and uses a search application. Our proposal uses an entertainment application which was inspired by an idea proposed in [35]. The computer game in this proposal is played by two unrelated players using the web. Whenever both players type the same keyword for the same image they win points given the fact that, the words that come from different people are more robust and descriptive than words typed individually. Our proposal is different because it uses machine learning techniques to improve the image annotation task.

3 System overview

In this section an overall description of the proposed system is presented. The system is composed of (see Fig. 1): a multimedia retrieval system, image search applications and a semi-automatic image annotation application.

The image retrieval method is based on image semantic analysis that uses multimodal information automatically extracted from images. Visual features, audio information and contextual metadata are explored to train semantic concepts that can be used directly for retrieval purposes by the image search applications (see Fig. 1) or to annotate picture with keywords in the semi-automatic application.

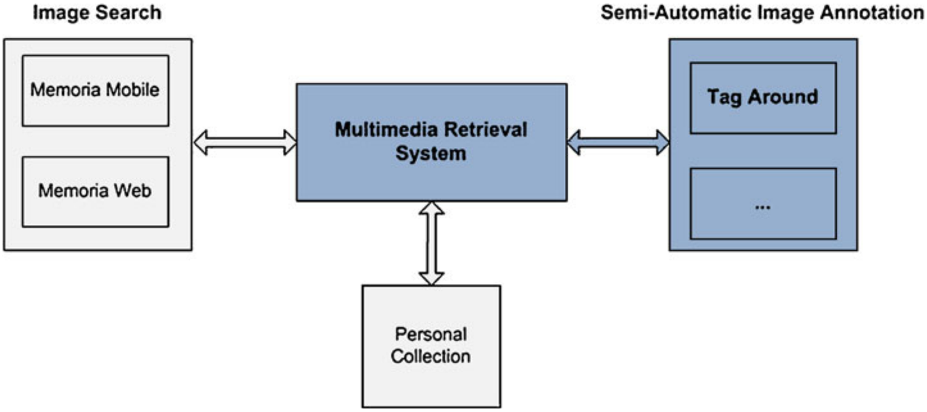


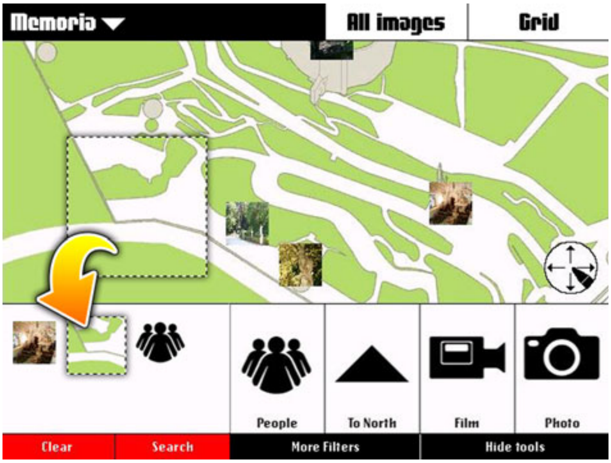
Fig. 1 System architecture

The image search applications are used to annotate images with audio information and contextual metadata (location and time) and to search pictures based on the proposed image retrieval system. Users can retrieve images using a mobile device (phone or PDA) when visiting a point of interest (see Fig. 2) or the Web before or after the visit.

Memoria Mobile (see Fig. 2) allows the retrieval of images of previous experiences during the visit and and Memoria Web enables the retrieval of the experiences before or after the visit. Both interfaces include a map of the place which can be used to define geographic queries and can help to guide the visit.

Memoria Mobile [13, 14], is an application to capture, share and access personal memories composed of pictures when visiting sites of interest. This application provides automatic image annotation at capture time using audio information and GPS data. Memoria Web [38] is an application developed to virtually visit the place and provide more annotations manually. Both interfaces include a map of the place

Fig. 2 Image search applications: Memoria mobile



which can be used to define geographic queries and can help to guide the visit. Both applications use the image retrieval system to access the memories.

The semantic models (image retrieval) are also included in a framework for semi-automatic image annotation. This platform combines the automatic method based on semantic concepts with the manual annotation through an application. This framework can include any type of application that follows a set of requirements. Our proposal instantiates the application block of the framework with a gesture based image annotation game.

The remainder of the paper describes the two darker blocks of Fig. 1, the multimedia information retrieval system based on semantic concepts and the application for semi-automatic image annotation.

4 Multimedia information retrieval system

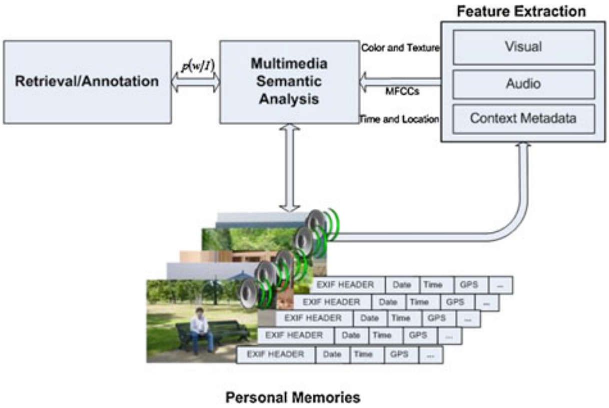
This section describes the methods to retrieve and annotate images based on semantic models trained with multimodal information. These methods correspond to three components (see Fig. 3):

- Image annotation and retrieval—application of the semantic analysis;
- Multimedia semantic analysis—estimation of the semantic models based on features extracted from images;
- Feature extraction—extraction of content and context information from images.

The first block applies the semantic models, represented by $p(w|I)$, to retrieve and to annotate images. Image retrieval is performed through semantic queries and picture annotation is obtained by associating semantic concepts to images as a way to describe their content.

In the multimedia semantic analysis block, the models are trained using visual and audio information and contextual metadata obtained at capture time. The method uses binary classification and a sigmoid function. The output of the block are the probabilities, $p(w|I)$.

Fig. 3 Block diagram of the proposed methods for annotation and retrieval



The techniques used to automatically extract this information from images are applied in the third block. It is assumed that each database document is composed of an image, an audio file and contextual metadata obtained at capture time.

The following subsections provide more details about these components. First is described our proposal for semantic image analysis. Then is explained how this method is used for image retrieval. The section ends with the methods used for features extraction.

4.1 Semantic analysis

The objective of the semantic analysis is to train a set of semantic models in order to obtain the probabilities $p(w|I)$ (see Fig. 3). To achieve this goal, the method described in this section uses visual and temporal information. Figure 3 shows that audio and spatial information is also extracted from images, but it is only used in the retrieval task as explained in Section 4.2. Concerning the audio information, the semantic analysis is performed using ASR (Automatic Speech Recognition) tools.

The proposed semantic description method is based on a combination of individual detectors. It uses a binary classifier to detect the presence or absence of a concept in an image and a sigmoid function to normalize the classifier output. After this, the temporal correlation between sequential images are analyzed to correct errors of the classification process.

4.1.1 Visual information

We use the Regularized Least Squares Classifier (RLSC) [29] as a binary classifier to detect a concept in an image. The output of the classifier can be used for image annotation, however this measure is not normalized therefore not suitable to combine different features or several concepts. The output of the classifier must be converted to a probability. We adapt the method proposed in [28] to the RLSC.

Assuming w is a Bernoulli random variable where the outcome can be one of two concepts (e.g., "Indoor"/"Outdoor", "Beach"/"No Beach" or "People"/"No People"), the probability $p(w|x)$ can be obtained using the output of the classifier $f(x)$ and a sigmoid function [28],

$$p(w|x) = \frac{1}{1 + e^{-Af(x)+B}}, \tag{1}$$

In [28] several methods to estimate the A and B parameters are discussed. Currently, we set them manually but in the future they will be estimated.

Given the training set $S_m = \{(x_i, y_i)_{i=1}^{m}\}$ where labels $y_i \in \{-1, 1\}$ and x_i is a vector of image features, the decision boundary between the two classes (e.g., "Indoor" and "Outdoor") is obtained by the discriminant function,

$$f(x) = \sum_{i=1}^{m} c_i K(x_i, x), \tag{2}$$

where $K(x_i, x)$ is the Gaussian Kernel $K(x_i, x) = e^{-\frac{\|x_i - x\|^2}{2\sigma^2}}$, m is the number of training points and $c = [c_1, ..., c_m]^T$, is a vector of coefficients estimated by Least Squares [29],

$$(m\gamma I + K)c = y, \tag{3}$$

where I is the identity matrix, K is a square positive definite matrix with the elements $K_{i,j} = K(x_i, x_j)$, y is a vector with coordinates y_i and γ is a regularization parameter. To choose the optimal values for σ and γ the cross-validation method is used.

A point x with $f(x) \leq 0$, is classified in the negative class $(y = -1)$, and a point with $f(x) > 0$ is classified in the positive class $(y = 1)$. If multiple features are used different classifiers are obtained.

4.1.2 Temporal information

Temporal proximity can improve the semantic analysis of images, captured by the same user, by including the temporal correlation in the model. For instance, Fig. 4 shows three sequential images captured with an interval of 10 s. It is not difficult to identity the concept "Beach" in Fig. 4a and c using visual information but the same does not happen in Fig. 4b. In this situation, the temporal proximity can help to infer the "Beach" concept in Fig. 4b, using the probabilities obtained by the other two.

Let $T = [t_1, t_2, ..., t_N]$ be the ranked vector with the capture time of each picture of a collection $C_{\text{img}} = \{I_1, I_2, ..., I_N\}$, the probability of a concept w given an image I_{t_i} taken at instant t_i is,

$$p_t(w|I_{t_i}) = \frac{\alpha_{t_{i-1}} p(w|I_{t_{i-1}}) + p(w|I_{i_t}) + \alpha_{t_i} p(w|I_{t_{i+1}})}{1 + \alpha_{t_{i-1}} + \alpha_{t_i}}, \tag{4}$$

where α_{t_i} and $\alpha_{t_{i-1}}$ are weights that measure the relevance of the $p(w|I)$ in the temporal adjacent images of the picture I_{t_i}. These weights are inversely proportional to the temporal distance between images,

$$\alpha_{t_i} = 1 - \frac{d(t_i)}{d_{\max}} \tag{5}$$

where $d_{\max}$ is a constant obtained empirically to represent the maximum time distance allowed to influence the adjacent images, $d(t_i)$ is the time distance between the captured image on instant t_i and the next one,

$$d(t_i) = \begin{cases} |t_{i+1} - t_i|, & |t_{i+1} - t_i| < d_{\max} \\ d_{\max}, & \text{other cases.} \end{cases} \tag{6}$$

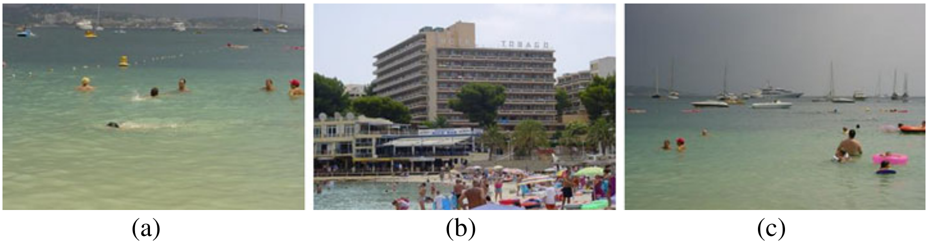


Fig. 4 Sequential images captured with a time interval of 10 s

This technique can be adapted to explore spatial correlations using the GPS coordinates of each image instead of the time information. Probabilities $p(w|I)$ are obtained using visual information.

4.2 Image retrieval

The goal of this component is to create a list of ranked images according to their relevance to the query. Considering a query defined by k concepts $Q = \{w_1, w_2, ..., w_k\}$ describing the background and some objects presented in the desired pictures (e.g., indoor, people and computers), the position in the ranked list of a picture I, using audio, visual, spatial and temporal information is obtained by the similarity measure,

$$Sim(Q, I) = f_{\text{gps}} \left[Sim_{\text{visual+time}}(Q, I) + Sim_{\text{audio}}(Q, I) \right], \qquad (7)$$

where f_{gps} is a filter applied when the query includes geographic elements (e.g., region or direction) to select images from the list obtained using the others components. The similarity using the audio information is defined by Sim_{audio} and $Sim_{\text{visual+time}}$ represents the similarity obtained using the visual and temporal information.

Considering multiple visual features, the visual and temporal similarity, $Sim_{\text{visual+time}}$, of an image I to a given query Q is the sum of the similarity obtained for each feature,

$$Sim_{\text{visual+time}}(Q, I) = \sum_{j=1}^{r} a_j Sim_{\text{visual+time}}(Q, x^j), \qquad (8)$$

where r is the feature number, x^j represents the j^{th} feature extracted from the I image and a_j is the weight of each feature assuming $\sum_{j=1}^{r} a_j = 1$. The ranked list obtained for each feature is given by the joint probability,

$$Sim_{\text{visual+time}}(Q, x^j) = p_t(w_1, w_2, ..., w_k|x^j), \qquad (9)$$

Assuming independence between concepts, the joint probability of a set of concepts given an image is,

$$p_t(w_1, w_2, ..., w_k|x) = \prod_{i=1}^{k} p_t(w_i|x^j), \qquad (10)$$

The probabilities $p_t(w_i|x^j)$ are computed in the semantic analysis section.

The function f_{gps} is used when the query includes geographic information. Two types of query can be used: region query and direction query. In the first case, images inside the defined region are selected. For each photo, Ig, the distance between its GPS location, and the location of the center of the selected region Qg is calculated using the Great Circle distance [32],

$$dist(Qg, Ig) = r_{\text{earth}} \Delta\phi \qquad (11)$$

where $r_{\text{earth}} = 6{,}378.7$ km is the earth radius and $\Delta\phi$ is,

$$\Delta\phi = \cos[\sin(lat_{Qg}) \sin(lat_{Ig}) + \cos(lat_{Qg}) \cos(lat_{Ig}) \cos(lon_{Ig} - lon_{Qg})] \qquad (12)$$

In this equation *lat* represents the latitude and *lon* the longitude. All the images satisfying the condition,

$$dist(Qg, Ig) < r_{\text{query}}, \tag{13}$$

are selected to be ranked. The r_{query} is the radius of the circle defined by the region selected.

The direction query is defined if the query includes the directions: North, South, East or West. Given a position in GPS coordinates and the direction, the system searches all of the pictures that are in the selected direction.

In both cases, if the query only contains geographic information, images are ranked according to the Great Circle distance (see (11)).

Recognized words from audio captured when the picture is taken are used to calculate the $Sim_{\text{audio}}(Q, I)$. This similarity is obtained using standard techniques of text retrieval [1]. Images are ranked according to this similarity measure.

4.3 Feature extraction

Audio information is converted to text using ASR tools. We used a speech recognition API developed by the Microsoft Language Development Center in Portugal. When the picture is captured, the camera parameters are saved in the EXIF header (a part of the JPEG header). These parameters include the capture time and the GPS coordinates. Concerning the visual features, four are used: (1) marginal HSV color moments, (2) Gabor filter features, (3) bags of SIFT descriptors and (4) bags of color regions (see [12, 13] for more details).

5 Semi-automatic image annotation

This section describes a framework for semi-automatic image annotation that uses the manual user intervention to correct errors of the automatic methods, in a way that is similar to a search with relevance feedback. The framework is composed of four parts (see Fig. 5): an application, the human interaction module, a block to update the automatic models and a method for automatic image annotation.

The application block can be a search application with relevance feedback or any other application where the user makes connections between images and words. To

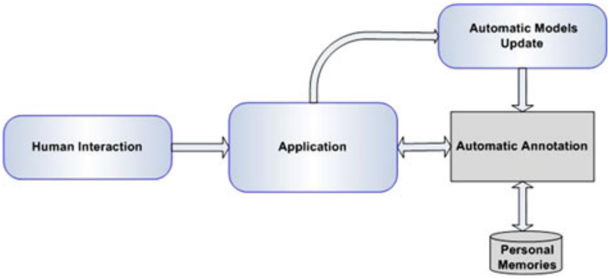


Fig. 5 Block diagram of the semi-automatic annotation method

motivate the user and inspired by the ESP GAME [35], a computer game called Tag Around was designed as the application module of the proposed framework. Section 5.1 describes the game, the human interaction module and the method to compute the score.

The automatic block uses the output of the semantic concepts (see Section 4.1) to automatically annotate images. The feedback given by the users is used to update the semantic concepts (Automatic models update block), that is, labeled images by users playing the game are included in the training set. Then, the semantic models are estimated again (see Section 4.1).

With the four blocks of Fig. 5, an algorithm for semi-automatic image annotation is defined. Section 5.2 describes this algorithm.

5.1 Tag Around

Our proposal for the application block of the framework is the Tag Around game [10, 15]. The goal of the application is to make annotations in a set of pictures. Figure 6 presents the interface of this game. The game is played with gesture input. A video camera is pointed to the user to capture the movements. A set of concepts is is displayed at top of the screen, grouped in a rotational platform, which is controlled by the user input. At the bottom of the screen the pictures are displayed, also controlled by the user. Whenever the users decide to match the center annotation/picture, a short animation sequence is presented to them, and a comment appears on the screen, which helps to know if a correct annotation was performed. To assist the user interaction while playing the game, the user picture is displayed with a set of hotspots.

While playing the game, the users have to correctly annotate the maximum number of images (in a given time period), using the words in the screen. Based on the user behavior (how well users tagged images) confidence is assigned to them. The game scoring is based on the automatic annotation algorithm, the trust in the user and the relevant feedback given by other users that have already played the game and tagged that set of pictures.

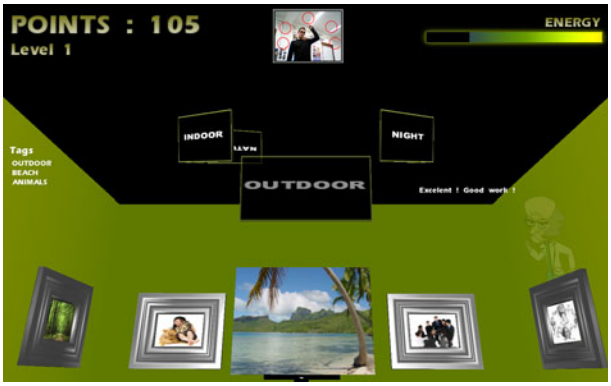


Fig. 6 Tag Around—a game for image tagging

5.1.1 Human interaction

As mentioned, a camera is pointed to the user to capture the movements. There are five hotspots in the image. The user has to be able to touch them to move the objects on the screen. There are two hotspots situated at the bottom of the screen (see Fig. 6) for picture rotation, and two hotspots situated at the top of the screen, to rotate the tags. The left side is used to rotate the picture left and the right hotspot is for the right rotation. The other two rotation hotspots are situated at the shoulders height, to handle tags rotation. The fifth hotspot is situated above the user and is used for matching purposes. Whenever users are sure of a picture-tag pair, they move their hand in that hotspot area to perform tag matching.

5.1.2 Score computation

When users tag an image the game module calculates the score of the player move, using a formula that includes the trust level in the player, the probability of a tag given an image (obtained by the automatic algorithm output) and the feedback provided by all previous users. If this result is a strong value, it is considered a correct annotation, and the user score and trust level are increased, increasing indirectly the feedback provided by all users.

Given a set of pictures $L = \{I_1, ..., I_{N_l}\}$ $(L \subset C_{\text{img}})$ and a set of concepts $V_{sc} = \{w_1, ..., w_{N_{\text{con}}}\}$ $(V_{sc} \subset V_{\text{con}})$, the score obtained by matching the concept w in the image I is computed by,

$$S_{\text{total}}(I, w, n, m) = C_{\text{group}}(m) + [1 - C_{\text{group}}(m)]S_{\text{new}}(I, w, n), \tag{14}$$

where n represents the number of correct annotations provided by the user, m is the number of times the concept w was annotated in image I, S_{new} is function that evaluates the annotation using the semantic concepts and the trust in the player (see (16)) and $C_{\text{group}}(m)$ means the group trust obtained by the correctness of the same annotation provided by other users,

$$C_{\text{group}}(m) = 1 - e^{-(\frac{m}{k_g})}, \tag{15}$$

the exponential parameter k_g is estimated in order to obtain a group trust near the maximum value after m annotations. We considered that three players providing the same annotation ($m = 3$) means an high group trust and for this reason k_g is obtained assuming this condition. The ESP GAME [35] validates an annotation with two players. With this equation, when $m = 2$, the score is not the maximum but is a value that allows the system to classify the annotation as correct.

When a concept w is annotated for the first time in an image I the score is computed by,

$$S_{\text{new}}(I, w, n) = C_{\text{player}}(n) + [1 - C_{\text{player}}(n)]p(w|I), \tag{16}$$

where $p(w|I)$ is the probability obtain by the automatic method (semantic concepts) and C_{player} is the trust of the system in the player that denotes the quality of previous annotations provided by the player,

$$C_{\text{player}}(n) = \begin{cases} k_p n, & n < K_{\text{moves}} \\ k_{\text{conf}}, & n \geq K_{\text{moves}} \end{cases} \tag{17}$$

where K_{moves} is a constant with the number of good moves to reach to the player trust maximum value k_{conf} and k_p is a constant that is used to increment the player trust.

The number of correct moves n increases when the group trust is different from zero and the score is greater than a defined threshold. It decreases when the score is above another threshold. These thresholds were obtained empirically. When the group trust is zero this means the score is obtained using only the semantic concepts and the player trust. In these cases, it is difficult to know the correctness of the annotation.

5.2 Image annotation algorithm

An annotation on an image $I \in C_{\text{img}}$ of a concept w_i belonging to a vocabulary $V_{\text{con}} = \{w_1, w_2, ..., w_k\}$, is defined as, $A(I, w_i)$. Given a set $L \subset C_{\text{img}}$ with N_l images and a set of $V_{sc} \subset V_{\text{con}}$ with N_{con} concepts, the semi-automatic method is defined by the following steps:

1. The subsets L and V_{sc} are presented in the interface (application block);
2. The user selects one image $I_l \in L$ and a concept $w_k \in V_{sc}$;
3. The user makes an annotation, $A_i(I_l, w_k)$;
4. The score is computed using the automatic models $p(w_k|I_l)$, the trust of the game in the player and the feedback provided by all previous users;
5. For all concepts $w_k \in V_{sc}$, if the $|\{A_1, A_2, ..., A_{N_A}\}| > N_{upd}$ for a concept w_k, then the training set is updated and the model for the concept w_k is computed again;
6. Go to 2.

A semantic model is trained again when the number of different correct annotations with the concept is above N_{upd}. An annotation is considered correct when it is performed by at least two users. As a result of this algorithm, a set of annotations $A = \{A_1, A_2, ..., A_{N_{\text{total}}}\}$ is obtained and the semantic concepts of the set V_{con} are estimated with a larger training set. If two different players provide the same wrong annotation the algorithm fails and this can increase the number of failures of the related concept but this is not a usual situation.

Both subsets, L and V_{sc}, used in each level of the game are selected in the automatic annotation block. Therefore, the learning process is driven by the automatic model.

6 Experiments and evaluation

To evaluate our proposal for image annotation, we start by testing the automatic method (based on semantic concepts) in images using different combinations of features and with users to assess the quality of the results. Then, we evaluate the concept detection including the feedback provided by the users playing the game. Finally, we evaluate the Tag Around game with usability tests since users are an important part of the process.

We train a set of concepts, suitable for personal collections, selected from the set of the 449 LSCOM [26] concepts. These concepts are estimated using a training set obtained from the Corel Stock Photo CDs, from the TRECVID2005 database and from Flickr, in order to build a generic data set. Nine semantic concepts were selected for evaluation: “People”, “Face”, “Outdoor”, “Indoor”, “Nature”, “Manmade”, “Snow”, “Beach” and “Party”.

6.1 Datasets

Our proposals were tested with two different picture collections. A personal collection with about 5,000 pictures was used to test the visual features and the time information. It was also used in the Tag Around game. This collection is essentially composed of pictures of people, nature or urban scenes, holidays and parties. We also tested the automatic method with a database of pictures taken by several visitors of a cultural heritage site in Sintra, Portugal. This place is composed of beautiful gardens, caves and romantic buildings. Given the nature of the place, only five concepts out of the nine were used in this evaluation.

6.2 Automatic method

This section starts by presenting the results obtained with the automatic detection of the semantic concepts in images with several visual features and time information. Table 2 compares the performance of the system using two different combinations of visual features and including time information in the second combination. Feature set 1 represents the combination of color moments with features obtained with the Gabor filters and Feature set 2 denotes the combination of a bag of color regions with a bag of SIFT descriptors. In this test we used the personal collection with 5,000 images.

Generally, the combinations that use bags are better than the other combinations of visual features. If time is included in the better combination, the system results improve. The concepts “Snow”, “Beach” and “Party”, exhibit the best increments when using time information. These concepts represent events where people stay during a larger period of time and for this reason, the probability of capturing correlated images increases. The concepts, “Outdoor” and “Snow”, present the best results with Feature set 1.

Table 2 Mean average precision (MAP) obtained for a set of concepts combining several visual features and time information. Maximum time distance considered between images was $d_{\max} = 240$ s

Concepts	Feature set 1	Feature set 2	Feature set 2 + time
People	0.69	0.75	0.75
Face	0.50	0.39	0.41
Outdoor	0.91	0.87	0.89
Indoor	0.59	0.57	0.60
Nature	0.45	0.57	0.58
Manmade	0.61	0.71	0.73
Snow	0.17	0.09	0.13
Beach	0.26	0.34	0.42
Party	0.14	0.22	0.26
MAP	0.48	0.50	0.53

Table 3 MAP obtained for a set of concepts using visual features, audio information and location data to select a region of 60 m

Concepts	Visual	Visual + audio	Visual + GPS	Visual + audio + GPS
Outdoor	0.86	0.86	0.97	0.97
Indoor	0.33	0.37	0.09	0.09
Nature	0.68	0.69	0.84	0.86
Manmade	0.32	0.70	0.68	0.71
People	0.23	0.27	0.20	0.20
Indoor + manmade	0.21	0.26	0.16	0.17
Outdoor + nature	0.75	0.75	0.86	0.84
MAP	0.48	0.56	0.54	0.55

To evaluate the inclusion of the audio information and the GPS data we used the database of a cultural heritage site in Sintra, Portugal. Table 3 presents the results obtained individually using visual features, audio information and the combination of the visual information with audio and GPS data. We tested the image retrieval method in several locations of the place. Table 3 presents the performance obtained in one of the tested locations. As shown, combining two types of data yields better results than using only visual features. The results obtained with the geographic metadata depend on the local features of the region selected, and for this reason, some concepts decrease their performance. For instance, the location selected is an outdoor region consequently, the concept "Indoor" obtained a low performance (next to last column in Table 3). Using the three types of information, because of the reason mentioned, including geographic metadata do not yields better results than using only visual and audio data.

To compare the results evaluated using the MAP measure with the users opinion about the same results, we ask to 58 voluntary participants to classify the results obtained by several searches using semantic concepts. The participants were graduated students. Ten of them were female. The users ranged in age from 21 to 31 years old with a mean age of 23.5.

After each search, users had to classify the results using a 5-point Likert-type scale, where 1 (one) means bad and 5 (five) means excellent. The users' classification is summarized in Table 4.

In general, the results obtained were reasonable for the users. This means that the values obtained using the MAP measure for the "People" and "Nature" concepts represent acceptable results for these users. In this experiment, we used the Features set 1 ("People", MAP = 69% and "Nature", MAP = 45%).

Table 4 Evaluation of the results obtained by several searches provided by 58 users

Queries	Mean	Standard deviation	Mode
People	3.9	1	3
Nature	3.8	1	4
Outdoor *and* no beach	4	1	4
People or nature	3.5	1	4

Table 5 MAP obtained using different training sets to estimate the semantic concepts: initial training set, with more 20 and 40 images for each concept

Concepts	Training set	Training set + 20	Training set + 40
People	0.75	0.81	0.82
Face	0.41	0.54	0.61
Outdoor	0.89	0.96	0.96
Indoor	0.60	0.78	0.80
Nature	0.58	0.80	0.83
Manmade	0.73	0.90	0.92
Snow	0.13	0.71	0.82
Beach	0.42	0.60	0.66
Party	0.26	0.44	0.54
MAP	0.53	0.73	0.77

6.3 Semi-automatic method

The section presents the results obtained by including users to correct the errors provided by the automatic method using a computer game. Table 5 presents the mean average precision (MAP) obtained using the initial training set and with more 20 and 40 images that were annotated using the game. Generally, best results occur when 20 new images are included in the training set. This happens because images of the test collection are included in the training set and the images of the same collection present more correlation. After that, when increasing the training set for each concept by 20 new images, the mean average precision obtained with the semantic models increases 4%.

The Tag Around game was subject to usability testing [11], with the goal to evaluate the interface complexity, the usefulness, the aesthetic aspects and to understand how easy it is to learn and use. These tests were performed by 15 voluntary participants ranging in age from 18 to 31 years old with a mean age of 24. After testing the application, users were asked to fill in a questionnaire and express their opinions regarding the application they have just tested. Table 6 shows the results obtained with usefulness related questions which are important to evaluate the proposal as an annotation tool. Questions presented were answered on a 5-point Likert-type scale, where 1 means totally disagree and 5 means totally agree. In general, the results are positive. The lack of consensus of the users about the use of the application to annotate their own images is a concern. Nevertheless, the feedback related with the use of the application as a game to spend time in a public space is positive, which was one of the intended design goals.

Table 6 Mean (μ) and standard deviation (σ) obtained for a set of usefulness related questions

Questions	μ	σ
“It was fun to use the application”	4.47	0.64
“I would use the application in a public place while waiting”	4.4	0.83
“I would use the application to have fun with family and friends”	4.5	0.74
“I would use the application to catalogue my own images”	3.53	0.99
“It would be more fun to annotate my own images with my own annotations”	3.93	1.03

7 Conclusions and future work

This paper proposes methods to annotate images using semantic concepts in two ways: automatically and in a semi-automatic framework. The semantic concepts are estimated using multimodal information. The semi-automatic image annotation framework is based on a computer game played with gesture input. In general, the results presented show that the semantic concepts increase the performance when including more types of data in the process. Time information improves the results of all concepts specially the concepts related to events with a certain duration. GPS data also improves the results but conditioned by the features of the selected place. Manual annotations are a good choice in terms of efficiency. However, users must be motivated to do it. Our solution to motivate the users is to convert the image annotation task in a enjoyable game. While people enjoy playing the game they contribute to solve the complex annotation problem. For future work, we are going to include more different data in the semantic model and we plan to develop an active learning method for the computer game.

References

1. Baeza-Yates R, Ribeiro-Neto B (1999) Modern information retrieval. Addison Wesley
2. Barnard K, Forsyth D (2001) Learning the semantics of words and pictures. In: International conference on computer vision, pp 408–415
3. Belkhatir M (2008) An agent framework based on signal concepts for highlighting the image semantic content. In: DEXA '08: proceedings of the 19th international conference on database and expert systems applications, pp 465–478
4. Chang E (2005) EXTENT: fusing context, content, and semantic ontology for photo annotation. In: CVDB '05: proceedings of the 2nd international workshop on computer vision meets databases, pp 5–11
5. Choi JY, Seungji R, Yong N, Plataniotis K (2008) Face annotation for personal photos using context-assisted face recognition. In: MIR '08: proceeding of the 1st ACM international conference on multimedia information retrieval, pp 44–51
6. Datta R, Joshi D, Li J, Wang J (2008) Image retrieval: ideas, influences, and trends of the new age. ACM Comput Surv 40(2). doi:10.1145/1348246.1348248
7. Duygulu P, Barnard K, Freitas J, Forsyth D (2002) Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Proceedings of the 7th European conference on computer vision. LNCS, vol 2353. Springer, pp 97–112
8. Fan J, Gao Y, Luo H (2007) Hierarchical classification for automatic image annotation.In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, pp 111–118
9. Frohlich D, Kuchinsky A, Pering C, Don A, Ariss S (2002) Requirements for photoware. In: Proceedings of the ACM conference on computer supported cooperative work, pp 166–175
10. Gonçalves D, Jesus R, Correia N (2008) A gesture based game for image tagging. In: CHI '08: CHI '08 extended abstracts on human factors in computing systems. ACM Press, New York, pp 2685–2690
11. Gonçalves D, Jesus R, Grangeiro F, Romão T, Correia N (2008) Tag around: a 3D gesture game for image annotation. In: ACM SIGCHI international conference on advances in computer entertainment technology (ACE 2008)
12. Jesus R, Abrantes A, Correia N (2006) Photo retrieval from personal memories using generic concepts. In: Advances in Multimedia Information Processing—PCM 2006. LNCS, vol 4261. Springer, pp 633–640
13. Jesus R, Dias R, Frias R, Abrantes A, Correia N (2007) Sharing personal experiences while navigating in physical spaces. In: 5th workshop on multimedia information retrieval. ACM SIGIR conference on research and development in information retrieval (SIGIR07)

14. Jesus R, Dias R, Frias R, Abrantes A, Correia N (2008) Memoria mobile: sharing pictures of a point of interest. In: Proceedings of the working conference on advanced visual interfaces (AVI '08). ACM, New York
15. Jesus R, Gonçalves D, Abrantes A, Correia N (2008) Playing games as a way to improve automatic image annotation. In: Proceedings of IEEE international workshop on semantic learning applications in multimedia (SLAM08). In conjuntion with CVPR08
16. Jiebo L, Boutell M, Brown C (2006) Pictures are not taken in a vacuum—an overview of exploiting context for semantic scene content understanding. IEEE Signal Process Mag 22:101–114
17. Jing F, Li M, Zhang H, Zhang B (2005) A unified framework for image retrieval using keyword and visual features. IEEE Trans Image Process 14(7):979–989
18. Kustanowitz J, Shneiderman B. Motivating annotation for personal digital photo libraries: lowering barriers while raising incentives. Technical report, HCIL, Univ. of Maryland
19. Lavrenko V, Manmatha R, Jeon J (2003) A model for learning the semantics of pictures. In: Neural information processing system conference
20. Lew M, Sebe N, Djeraba C, Jain R (2006) Content-based multimedia information retrieval: state-of-the-art and challenges. In: ACM transactions on multimedia computing, communication, and applications, pp 1–19
21. Li J, Wang J (2006) Real-time computerized annotation of pictures. In: ACM international conference on multimedia, pp 911–920
22. Lu Y, Zhang H, Wenyin L, Hu C (2003) Joint semantics and feature based image retrieval using relevance feedback. IEEE Trans Multimedia 5(3):339–347
23. Monay F, Gatica-Perez D (2003) On image auto-annotation with latent space models. In: Proceedings of the eleventh ACM international conference on multimedia, pp 275–278
24. Mori Y, Takahashi H, Oka R (1999) Image-to-word transformation based on dividing and vector quantizing images with words. In: Proceedings of the international workshop on multimedia intelligent storage and retrieval management
25. Naphade M, Huang T (2001) A probabilistic framework for semantic video indexing, filtering, and retrieval. IEEE Trans Multimedia 3:141–151
26. Naphade M, Smith J, Tesic J, Chang S, Hsu W, Kennedy L, Hauptmann A, Curtis J (2006) Large-scale concept ontology for multimedia. IEEE Multimed 13(3):86–91
27. Over P, Ianeva T, Kraaij W, Smeaton A (2006) Trecvid 2006 overview. NIST TRECVID-2006
28. Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Advances in large margin classifiers. MIT Press, pp 61–74
29. Poggio T, Smale S (2003) The mathematics of learning: dealing with data. In: Notice of American Mathematical Society, pp 537–544
30. Rodden K, Wood K (2003) How do people manage their digital photographs? In: Conference on human factors in computing systems (CHI 2003), pp 409–416
31. Shneiderman B, Kang H (2000) Direct annotation: a drag-and-drop strategy for labeling photos. In: Proceedings international conference information visualization (IV2000), pp 88–95
32. Sinnott R (1984) Virtues of the Haversine. Sky Telesc 68:158
33. Smeulders A, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
34. Tulving E (2002) EPISODIC MEMORY: from mind to brain. Annu Rev Psychol 53:1–25
35. von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: Proceedings of the SIGCHI conference on human factors in computing systems CHI '04, pp 319–326
36. Wenyin L, Dumais S, Sun Y, Zhang H, Czerwinski M, Field B (2001) Semi-automatic image annotation. In: Human–computer interaction—interact '01
37. Yan R, Natsev A, Campbell M (2007) An efficient manual image annotation approach based on tagging and browsing. In: ACM international workshop on the many faces of multimedia semantics, co-located with ACM multimedia, pp 13–20
38. http://di205.di.fct.unl.pt/instory-web. Last accessed: 15 Jan 2009

Rui Jesus got his MSc in Instituto Superior Técnico (IST/UTL) while he was a researcher at CEDET-ISEL working on articulated models for human motion tracking. During the summer of 2005 he was working at Imperial College London, in collaboration with the Multimedia Information Retrieval group where he participated in TRECVID 2005. Currently he is a Teaching Assistant at the Instituto Superior de Engenharia de Lisboa (ISEL), where he teaches courses on Digital Signal Processing and Information theory and he is member of a research group (Multimedia and Machine Learning) doing work on multimedia information retrieval and signal processing. He received the PhD degree from the New University of Lisbon, Portugal, in 2010. Currently he is also a researcher in the Interactive Multimedia Group of CITI, New University of Lisbon.

Arnaldo J. Abrantes received the Ph.D. degree from the Technical University of Lisbon, Portugal, in 1998. He is Associate Professor with the Electronics, Telecommunications and Computer Engineering Department of Instituto Superior de Engenharia de Lisboa (ISEL). His research interests are in computer vision, multimedia information retrieval and machine learning. He develops his activities at ISEL, within the group of Multimedia and Machine Learning.

Nuno Correia is a Professor at the New University of Lisbon, where he teaches Multimedia Computing and Image Processing, and heads a research group on multimedia information processing and interaction. He was a researcher at Interval Research, Palo Alto, CA and a researcher at INESC, Portugal. He participated in several EU funded projects, including Euromath, MADE that provided the foundation for the international standard PREMO (Programming Environment for Multimedia Objects) and D-ARTS (DVD Authoring Tools) and was Portuguese representative in standardization meetings. He has worked and directed projects on augmented environments and mobile storytelling funded by the Portuguese Science Foundation and on multimedia for learning funded by HP. Nuno Correia is co-director of the national program on Digital Media in cooperation with UT Austin and associate editor of the Computers and Graphics journal published by Elsevier. URL: http://img.di.fct.unl.pt/~nmc.