Eurographics Italian Chapter Conference (2010)
E. Puppo, A. Brogni, and L. De Floriani (Editors)

Toward Wide-Area Camera Localization for Mixed Reality

Valeria Garro and Andrea Fusiello

Department of Computer Science, University of Verona
Strada Le Grazie 15, 37134 Verona (Italy)

Abstract

This paper describes a work in progress towards the implementation of a complete system that provides tourists with relevant visual information related to cultural heritage sites. Thanks to the diffusion of high-end mobile devices and the recent improvement in computer vision research on 3D Structure and Motion reconstruction, it is now possible to develop mobile mixed reality applications that can interact with spots of historical interest in the city. In particular we present an accurate localization of the mobile device that leverages on a pre-computed 3D structure to obtain image-model correspondences. Preliminary experiments with a calibrated camera – indoor and outdoor – demonstrate sufficient accuracy to support mixed reality.

Categories and Subject Descriptors (according to ACM CCS): H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities I.4.9 [Image Processing and Computer Vision]: Applications—I.2.10 [Computing Methodologies]: Vision and Scene Understanding—

1. Introduction

The increasing popularity of new generation smartphones combined with recent results on 3D reconstruction from Structure and Motion (SaM) allows the design of fully autonomous system to support mixed reality applications. In particular in this paper we will describe our progress toward the development of system that assists tourists visiting cultural heritage sites and historical town centres. The proposed procedure provides to the user contextual video, text information and also 3D representations of particular points of interest of the city like buildings or monuments that can be visualized directly onto the image captured by the device's camera as a mixed reality display.

Nowadays robust SaM algorithms [SSS06], [FFG09] are available, from which we can obtain a sparse 3D reconstruction of the most interesting locations of the city, like squares, monuments or even entire quarters. This 3D model is usually represented as a 3D points cloud, as shown in Fig.1. Beyond spatial 3D information each of these 3D points is related also to a set of appearance descriptors like SURF [BETVG08] or SIFT [Low04], invariant to similarity transformation and robust to viewpoint and illumination changes. These descriptor will be our anchor bolt for the localization of a smarthphone's camera with respect to the 3D reconstruction coordinate frame. If the camera pose estimation reaches a good precision it becomes possible not only to roughly localize the user but also to render a graphic layer on the device's display where 3D models will interact coherently with the actual frame.

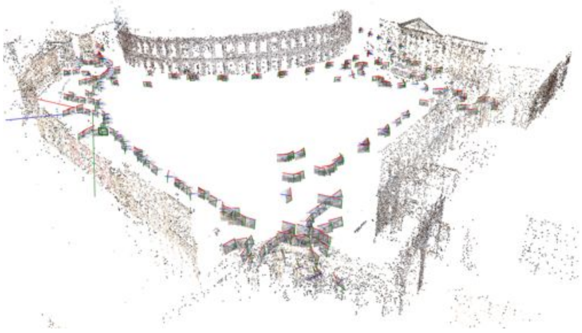


Figure 1: *3D reconstruction of Piazza Brà*

Basically, this system can provide a cultural knowledge that overcome the simple unguided tourist experience. In fact with this framework each user will have at his disposal an interactive guide when he is visiting a particular city just looking at monuments or historical buildings through his mobile device's camera. It will provide a personal guide to each potential user, improving access to cultural knowledge that otherwise could be more difficult to achieve, especially in an outdoor environment.

Despite the fact that other type of sensors are available like GPS and Wifi signals, it's still crucial adopting image-based techniques in order to obtain such an accurate position and orientation of the camera needed for applications illustrated above. One of the main reasons is the low-level accuracy provided by the other sensors when they are available. GPS accuracy is affected by atmospheric conditions, natural and artificial barriers and the error measured is usually between 1 to 10 meters. Furthermore the signal is completely absent in indoor environment. With this type of sensors we can not guarantee either a precise position or the orientation of the device and this is not sufficient to ensure an high level of correlation between the augmented reality layer and the framing area.

In literature, several systems (e.g. [RS04], [SB07]) have been presented providing the user additional information about the environment. These works describe first examples of tourist guide applications suitable for mobile devices. At the present time these approaches present a reductive functionality due to the complex data acquisition for the environment models and also for the hardware components employed. Remarkable results on mobile camera pose estimation have been reached also for indoor environment [HB08]. This approach is slightly different from ours; they use different feature detection method based on image segmentation exploiting the fact that their application is suitable for indoor localization. Due to recent smartphones increasing of hardware capabilities (e.g. the integration of high-resolution cameras and GPU processors) it is now possible to support also a real time mobile implementation of the most efficient features and descriptor such as SURF [CXG*07] and SIFT [WRM*08].

In this paper we will focus on accurate pose estimation algorithm of a camera with respect to a given 3D points cloud reconstruction. The rest of the paper is organized as follows. Section 2 describes our system in details: a pre-processing step in which we configure the dataset of available images adopting an efficient Bag-of-Visual-Words model and the "online" procedure of frame localization. Experimental results are shown in Section 3 and conclusions and further directions are presented in Section 4.

2. System description

Our system leverages on a SaM pipeline as baseline technology. This pipeline allows to produce a sparse set of 3D points endowed with appearance descriptors by processing a large set of images of the scene.

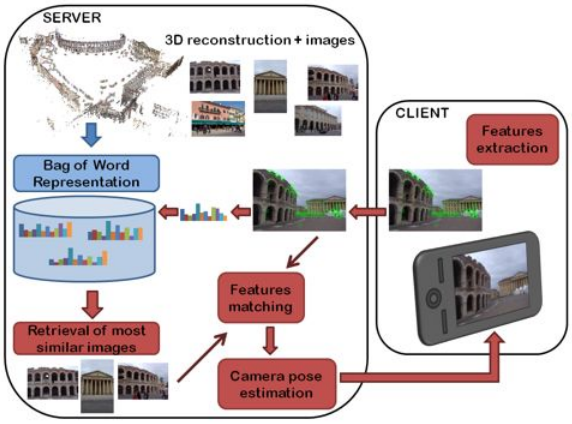

SERVER
3D reconstruction + images
CLIENT
Features extraction
Bag of Word Representation
Features matching
Retrieval of most similar images
Camera pose estimation

Figure 2: *System Overview. The offline data pre-processing is marked in blue, the online steps are drawn in red.*

In order to achieve a fast and precise localization of a mobile camera it would impractical and also ineffective to match the present image against the entire set of images. Instead we limit the matching to the subset of most similar images. This problem can be solved with image-based techniques for object recognition and scene classification [SZ03], [YJHN07] that provide an efficient image representation drawing inspiration from the text retrieval community. These methods, based on Bag-of-Words model, introduce the concept of "visual words" in analogy to representative words in text document. The local features (e.g. SIFT, SURF) extracted from an image correspond to single words in a text. Our approach differs slightly from the previous cited works, in addition to the set of images of the site it makes use also of the 3D reconstruction given by the SaM pipeline. The main idea is that using 3D points clouds model we have additional information with respect to image-based methods. The localization procedure not only relies on local features' correspondences between images but also it takes advantage of further geometric constraints given by the 3D data recovered. A similar approach is presented in [IZFB09].

Our system involves two main stages: the retrieval phase to determine a subset of images that are most similar to the query image and the computation of the camera's pose. This section describes the main steps of the proposed procedure: first an "offline" pre-computation stage is needed in order to set the Bag-of-Words retrieval model; then during the "online" stage the mobile camera's mobile camera's images are processed to calculate the location of the device.

Fig.2 illustrates the major stages of our system. A client-server architecture will enable applications based on large-scale database. In order to provide an efficient distribution of computational payload the mobile client device will perform only features extraction on the current image to localize. The data will be transmitted over an Internet connection to a server that will operate the localization procedure and send back contextual visual information to the smartphone.

2.1. Offline data pre-processing

The Bag-of-Words framework allows compact image representations and a scalable method for large image retrieval databases. As described above, the points cloud 3D model obtained by the SaM pipeline is the basis for the localization procedure together with the set of registered images. Each 3D point of the model is related to a set of SURF descriptors [BETVG08] derived from the corresponding features in the registered images. This stage aims to build the visual words codebook from this set of registered images to perform a efficient retrieval during the online phase. We adopt SURF instead of SIFT descriptor heading for high computing speed and low amount of memory storage in order to achieve real time performances.

The first step, which provides the construction of visual word codebook, consists in the quantization of the descriptors associated to the 3D points in order to keep a compact set of representative features (i.e. the clusters' centers), called "visual words". In literature several feature quantization approaches have been proposed: for a small database like ours a simple clustering technique like k-means in Euclidean space can be sufficient, however when the database size increases a more complex data structure must be employed which supports both a more compact representation of visual words and a more efficient search procedure. Two examples of this advance approach are vocabulary tree [NS06], that uses hierarchical k-means to recursively subdivide the feature space, and random forest [PCI*07].

The size of the vocabulary is one of the crucial points of the model. A coarse clustering can not be enough discriminative since two features descriptors with low similarity value can be assigned to the same visual word. In presence of big datasets, the number of elements to classify can be too large. In this case the clustering can be performed on a smaller subset of feature descriptors related to the most representative images carefully chosen from the entire dataset.

A second step computes for each image a compact representation as the histogram of occurrences of each visual word in the image. This representation is referred to as "BoW signature". It is customary to apply a weighting scheme to BoW signatures that considers visual words' frequencies both in a single image and in the entire database. The rational is that some visual words can be less distinctive due to a high frequency of appearance in the entire image database, and these items must be down-weighted; on the other hand, visual words appearing only in few images have a high distinctive power and should be up-weight. A weighting scheme commonly used in text retrieval is known as "term frequency-inverse document frequency" (TF-IDF). The TF-IDF scheme works as follows, given a visual word (term) t in an image (document) d its weight is given by:

$$\text{tf-idf}_{t,d} = tf_{t,d} \times idf_t. \tag{1}$$

The Term Frequency is defined as:

$$tf_{t,d} = \frac{n_{t,d}}{\sum_k n_{k,d}}. \tag{2}$$

where $n_{t,d}$ is the number of occurrences of the visual word t in the image d, and the denominator is the sum of occurrences of all terms in the image d. The Inverse Document Frequency is related to each visual word considering the entire database of images and is defined as:

$$idf_t = \log \frac{|D|}{1 + |\{i : n_{t,i} \neq 0\}|}. \tag{3}$$

where D is the set of all images and $\{i : n_{t,i} \neq 0\}$ is the set of images where the visual word t appears at least one time. We apply the TF-IDF weighting to each BoW signature of the database.

2.2. Online camera pose estimation

In the following we explain the online phase of our system. We first extract SURF features from the query image, then each feature is assigned to a visual word of the codebook using an efficient kd-tree structure and the BoW signature of the query is computed. Then we calculate the similarity between query and database images using the cosine similarity measure of the related BoW signatures, defined as:

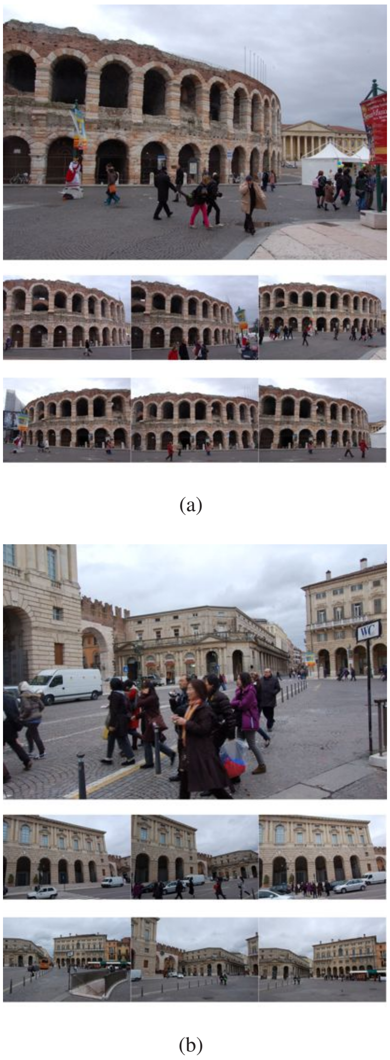
$$sim(BoW_q, BoW_i) = \frac{BoW_q \cdot BoW_i}{\|BoW_q\| \, \|BoW_i\|}. \tag{4}$$

for each image i belonging to the database D. A subset $\tilde{D}$ of m most similar images can now be determined.

The second step consists on selecting the SURF features associated to the points of the 3D model visible from the images in $\tilde{D}$. As a further additional constraint we choose only the features related to 3D points that are visible from more than one view. In this way we can discharge possible wrong retrieval results. We perform matching between the SURF features extracted from the query image and the SURF features just selected, obtaining a set of correspondences between 2D query points and 3D model points. Camera pose estimation can now be computed applying Fiore's linear algorithm [Fio01], if the intrinsic parameter of the mobile camera are known, or linear resection [HZ03] in case of uncalibrated device. In order to cope with possible outliers we use MSAC [TZ00]. A further refinement of camera pose can be done applying a non-linear refinement that minimizes the reprojection error of the set of 3D points inliers ensued from MSAC. The minimization is performed using the Levenberg-Marquardt algorithm.

3. Results

We first tested the performance of our system on a dataset of 318 images of "Piazza Brà" square in Verona of which we have a full 3D reconstruction provided by the SaM pipeline [FFG09], shown in Fig.1. The 3D points cloud is endowed with 200000 SURF descriptor, this means that for each image a set of 600 features on average is been matched and related to 3D points of the model. The resolution of all images is 1504 × 1000. We performed our test employing a leave-one-out approach. Each registered camera has been first removed from the dataset together with the related feature descriptors and then the localization algorithm has been run on the updated dataset. In this way we can consider the registered camera obtained with the SaM pipeline as our ground-truth data.


(a)

(b)

Figure 3: *Examples of retrieval. Top: the image to localize; bottom rows: the subset of the most similar 6 images.*

For the retrieval step, we have performed a k-means clustering on the SURF descriptors related a specific subset of 135 images (with $k = 10000$). By employing only a part of the datasets images we have improved the discriminative power of each visual word and at the same time we could increase the number of clusters computed. Two examples of the retrieval results are shown in Fig.3. In particular Fig.3b illustrates the robustness to occlusion provided by the Bag-of-Words approach. In Tab. 1 is shown the accuracy of our location algorithm in terms of Euclidean distance of the camera centre with respect to the ground truth data and the residual rotation angle given by the geodesic distance in $SO(3)$:

$$d_g(R_{gt}R_l) = \min\left\{\left\|\log R_{gt}^T R_l\right\|_F, \left\|\log R_l^T R_{gt}\right\|_F\right\}. \qquad (5)$$

where R_{gt} is the rotation component of the camera matrix of the ground truth data $P_{gt} = K_{gt}\,[R_{gt}|T_{gt}]$ and R_l is the rotation component of the camera matrix $P_l = K_l\,[R_l|T_l]$ computed by our algorithm.

Table 1: *Camera Pose Average Error*

Method	Camera Centre Distance [m]	Residual Rotation Angle [deg]
Fiore	0.2509	0.56
Resection	3.0101	4.03
Fiore + refin	0.1270	0.29
Resection + refin	3.0022	4.00

As an example of the mixed reality display that our system will provide, Fig. 4 illustrates the overlap of a graphic rendering of a the 3D model of the Arena onto the actual image. The visual alignment depends on the correct localization of the image with respect to the 3D model. In this example we used a calibrated image. Indeed, as it can be noticed from Tab. 1, while the accuracy in the calibrated case is fairly good, in the uncalibrated case (resection) the average error is still too high to enable a mixed reality display.

A second experiment has been run testing an indoor environment. The dataset is composed by 213 images (1504 × 1000) of a train station entrance hall. The 3D model obtained by the SaM pipeline is formed by 8650 3D points endowed with 28500 SURF descriptors. In this case the k-means clustering has been processed with $k = 5000$. For both experiments feature points have been extracted with 'Fast-Hessian' detector [BETVG08] setting the threshold equal to 300. In the matching stage we have adopted the evaluation criterion presented by Lowe [Low04] with a matching threshold of 0.5. Table 2 reports the related camera pose average error.

4. Conclusions and Future Works

We have presented a system architecture that provides a personal guide to each potential user that owns a recent mobile device when visiting indoor and outdoor cultural heritage sites. We have focused our paper on the development of an accurate camera pose estimation algorithm with respect to a

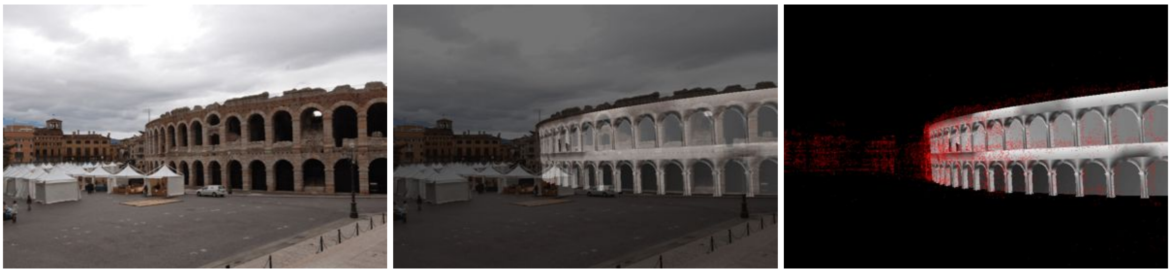
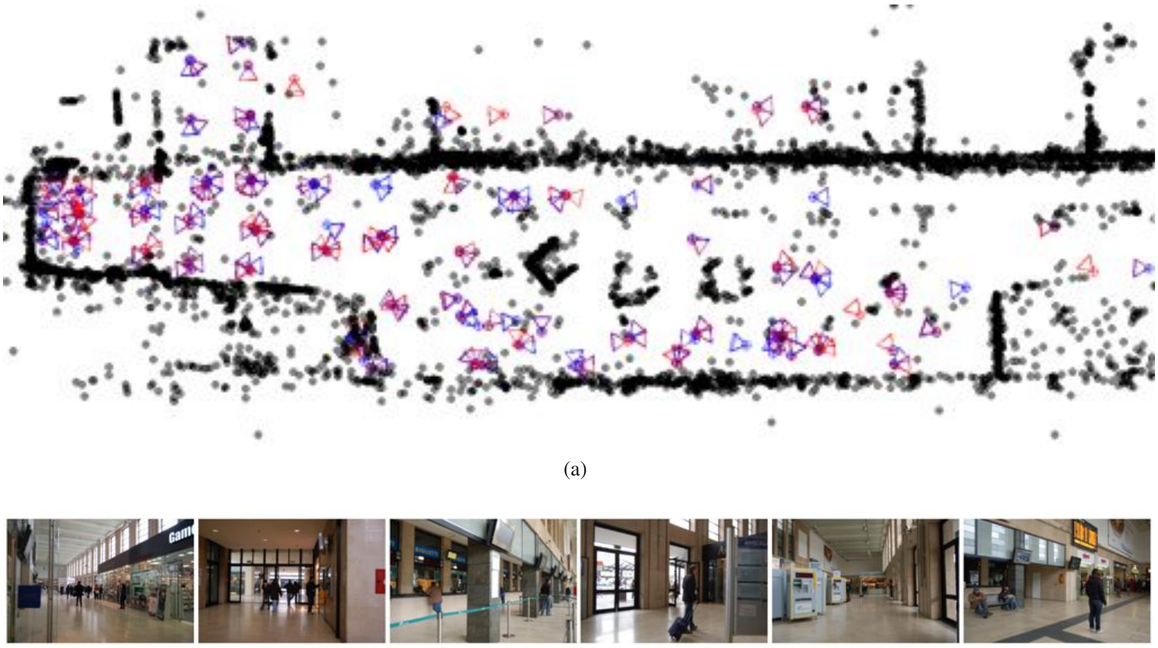


Figure 4: *Example of mixed reality overlay from a localized view. On the left the original image; in the middle a superimposed 3D model of the Arena; on the right the 3D model with the points cloud in red.*



(a)

(b)

Figure 5: *(a) Train station entrance hall 3D reconstruction. Blue cameras are the ground truth data; Localized cameras are marked in red. (b) a subset of images from the dataset*

Table 2: *Camera Pose Average Error*

Method	Camera Centre Distance [m]	Residual Rotation Angle [deg]
Fiore	1.8195	18.94
Resection	4.2765	33.14
Fiore + refin	0.1406	1.10
Resection + refin	2.2973	13.81

3D model given by a SaM pipeline. We have tested our system obtaining good results in the calibrated case. Based on these results our next efforts will concentrate on achieving comparable precision also with uncalibrated camera and on the real time mobile implementation of features extraction on the client side.

5. Acknowledgements

This work has been funded by the EU SAMURAI Project. The authors would like to thank RFI represented by Ing. Stefano de Muro for allowing the authors to take pictures the train station of Padova.

References

[BETVG08] BAY H., ESS A., TUYTELAARS T., VAN GOOL L.: Speeded-up robust features (surf). *Comput. Vis. Image Underst. 110*, 3 (2008), 346–359. 1, 3, 4

[CXG*07] CHEN W.-C., XIONG Y., GAO J., GELFAND N., GRZESZCZUK R.: Efficient extraction of robust image features on mobile devices. In *ISMAR '07: Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality* (Washington, DC, USA, 2007), IEEE Computer Society, pp. 1–2. 2

[FFG09] FARENZENA M., FUSIELLO A., GHERARDI R.: Structure-and-motion pipeline on a hierarchical cluster tree. In *Proceedings of the IEEE International Workshop on 3-D Digital Imaging and Modeling, ICCV Workshops* (Kyoto, Japan, 2009), pp. 1489–1496. 1, 3

[Fio01] FIORE P. D.: Efficient linear solution of exterior orientation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*, 2 (2001), 140–148. 3

[HB08] HILE H., BORRIELLO G.: Positioning and orientation in indoor environments using camera phones. *IEEE Computer Graphics and Applications 28* (2008), 32–39. 2

[HZ03] HARTLEY R., ZISSERMAN A.: *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2003. 3

[IZFB09] IRSCHARA A., ZACH C., FRAHM J., BISCHOF H.: From structure-from-motion point clouds to fast location recognition. pp. 2599–2606. 2

[Low04] LOWE D. G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision 60*, 2 (2004), 91–110. 1, 4

[NS06] NISTER D., STEWENIUS H.: Scalable recognition with a vocabulary tree. In *CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Washington, DC, USA, 2006), IEEE Computer Society, pp. 2161–2168. 3

[PCI*07] PHILBIN J., CHUM O., ISARD M., SIVIC J., ZISSERMAN A.: Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2007). 3

[RS04] REITMAYR G., SCHMALSTIEG D.: Collaborative augmented reality for outdoor navigation and information browsing. In *In Proceedings of the Symposium on Location Based Services and TeleCartography* (2004), Wiley, pp. 31–41. 2

[SB07] SCHMEIL A., BROLL W.: Mara - a mobile augmented reality-based virtual assistant. *Virtual Reality Conference, IEEE 0* (2007), 267–270. 2

[SSS06] SNAVELY N., SEITZ S. M., SZELISKI R.: Photo tourism: exploring photo collections in 3d. In *SIGGRAPH: International Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2006), pp. 835–846. 1

[SZ03] SIVIC J., ZISSERMAN A.: Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision* (Oct. 2003), vol. 2, pp. 1470–1477. 2

[TZ00] TORR P. H. S., ZISSERMAN A.: MLESAC: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding 78* (2000), 2000. 3

[WRM*08] WAGNER D., REITMAYR G., MULLONI A., DRUMMOND T., SCHMALSTIEG D.: Pose tracking from natural features on mobile phones. In *ISMAR '08: Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality* (Washington, DC, USA, 2008), IEEE Computer Society, pp. 125–134. 2

[YJHN07] YANG J., JIANG Y.-G., HAUPTMANN A. G., NGO C.-W.: Evaluating bag-of-visual-words representations in scene classification. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval* (New York, NY, USA, 2007), ACM, pp. 197–206. 2