DUPLICATE DETECTION IN CONSUMER PHOTOGRAPHY AND NEWS VIDEO

Alejandro Jaimes
Electrical Engineering Department
Columbia University
New York, NY 10027
ajaimes@ee.columbia.edu

Shih-Fu Chang
Electrical Engineering Department
Columbia University
New York, NY 10027
sfchang@ee.columbia.edu

Alexander C. Loui
Imaging Science and Technology Lab
Eastman Kodak Company
Rochester, NY 14650-1816
alexander.loui@kodak.com

ABSTRACT
Consumers often make more than one photograph of the same scene, creating non-identical duplicates and near duplicates. In Kodak's consumer photography database, on average, 19% of the images, per roll, fall into this category. Automatic detection of duplicates, therefore, is extremely useful in applications that help users organize their image collections. We introduce the challenging problem of non-identical duplicate image detection in consumer photography, describe *STELLA* (a novel interactive personal image collection organization system), and give an overview of our novel framework for detecting duplicate and near duplicate consumer photographs and news videos.

Keywords
Consumer photography, copy detection, digital album.

1. INTRODUCTION
More than ever, consumers are able to easily obtain digital versions of their photographs, build, and maintain digital image collections at home. In consumer photography it is common to have one or more non-identical photographs of the same scene. In Kodak's consumer image database [7], on average, 19% of the images, per roll, are either duplicates or near duplicates. In TV news broadcast it is common for videos of the same scene to repeat, extensively, on different channels and/or at different times.

Accurate and efficient detection of duplicates and near duplicates plays a fundamental role in several important applications (e.g., media tracking [3], copyright infringement detection [1], integrity in databases , security, filtering, etc.). In consumer photography detection of duplicates is crucial in developing applications that help users effectively organize their personal collections (e.g., [5][7][9][10]). In news video, duplicate detection can be used to generate news summaries for a time period (e.g., a day) to include only non-duplicate videos. Finding how the same video or image is used by different sources (e.g., broadcasters in different countries) can also have several interesting applications related to intelligence information gathering. Different audio usually accompanies repeated videos, so that visual content has to be used for duplicate detection.

Detection of non-identical duplicates has not been addressed in previous systems (e.g., [5][7][9][10]). Traditional approaches to measure similarity (e.g., based on global color histograms) are unsuitable for distinguishing between non-identical duplicates and very similar non-duplicates. Previous work on duplicate detection has focused only on images with minor variations (no camera/scene changes [3][1]). Although the work in [8] aims to cluster images based on similar views of the same scene, the specific duplicate problem is not addressed (e.g., no analysis of *differences* between duplicate candidates, among others).

2. STELLA
STELLA (Figure 1) is a system that helps users semi-automatically organize their images, for archiving as well as for producing digital albums. Photographs made with standard film cameras are scanned and input into the system. Then the images are automatically organized using a novel extension to Ward's clustering algorithm. Images are clustered hierarchically based on visual content and roll of film location (the source of the images is film, so time stamp information is not used and no metadata is available). The results of the clustering are presented to the user who can subjectively manipulate the clusters to organize his personal collection. A first needed step in the clustering process is the detection of duplicate and near duplicate images, since they form the first and most obvious level of clustering.

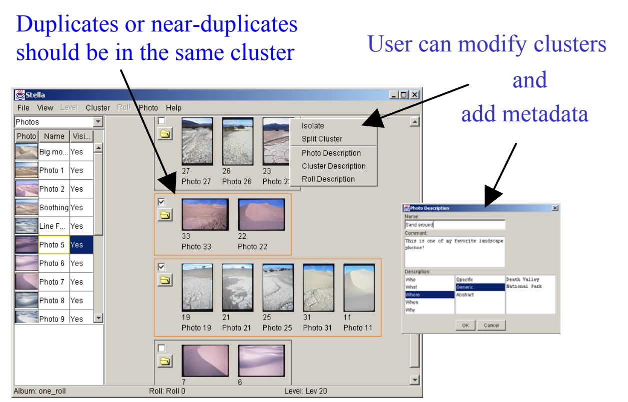

Figure 1. STELLA Graphical User Interface.

ACM Multimedia '02, December 1-6, 2002, Juan Les Pines, France.

3. THE DUPLICATE PROBLEM

An image is a duplicate of another, if it *looks* the same, corresponds to approximately the same scene, and does not contain *new* and *important* information. Two images (i_1, i_2), therefore, *do not have to be identical* (i.e., pixel by pixel) to be considered duplicates– whether two images are duplicates or not depends entirely on the *differences* between them. Differences can be accounted for due to changes in *subject*, *camera parameters*, the *scene*, or the *image*.

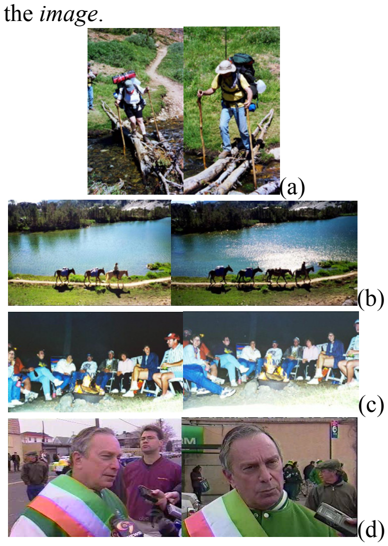

Figure 2. Example consumer duplicate candidates (a)-(c) and news video duplicates (d).

Automatically detecting duplicate images is extremely challenging. First, there is significant subjectivity in deciding if two images are duplicates. In an image duplicate database we have constructed [4] (255 image pairs; 60 rolls from 54 real consumers labeled by 10 other people) we found only 43% full agreement on duplicate/non-duplicate labels. Often, when there is agreement in the labels, we either find that high-level semantic information is used to make a decision, or the changes in visual similarity are too subtle (in non-duplicate cases) or too great (in duplicate cases). Therefore, two duplicate images can be significantly different visually, while two non-duplicate images can look almost the same.

4. A NOVEL DUPLICATE FRAMEWORK

We model duplicates in terms of the components (*scene*, *camera*, and *image*) that cause changes between two photographs of approximately the same scene and build a taxonomy of different types of duplicates. Our framework [4] consists of three stages: *(1) global coordinate transformation compensation; (2) detection of change areas; (3) analysis of change areas*. In the first and second stages of the framework we use low-level semantic information (i.e., from applying classifiers for simple object areas such as skin and sky) and the geometric properties of multiple images of the same scene.

We compute interest points, use a block-based correlation approach, and incorporate a novel image area saliency measure in the computations. The duplicate decision is largely based on the types of images that are being compared (e.g., similarity between two landscape images is not the same as between two group portraits). Therefore, in the third stage we analyze the change areas using class-specific similarity metrics using three distinct approaches: (1) direct comparison of visual features extracted from the two images; (2) classification of changes into a limited set of object *areas* (e.g., vegetation, sky); (3) application of specialized *object* detectors (e.g., face).

Finally, detectors for image areas (e.g., skin, grass, sky, etc.) are learned from training using the Visual Apprentice ([4]) and the final duplicate decision is made through statistical inference based on the outputs of the detectors, global image classifiers, and class-specific similarity metrics (Figure 3).

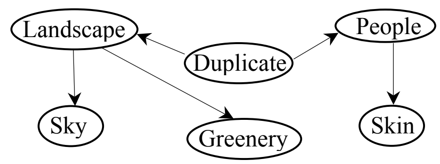

Figure 3. An example of the relationships in our graphical model for duplicate detection.

5. CONCLUSIONS

Duplicate detection is an important but very challenging problem. In this paper we have introduced the non-identical duplicate image detection problem and discussed *STELLA*, a system for helping users organize their personal image collections. We have also discussed a novel duplicate detection framework that is based on a model of the differences between duplicate image candidates. Our framework uses low-level semantic information and the geometric properties of multiple images of the same scene.

6. REFERENCES

[1] E.Y. Chang, et al., "RIME: A Replicated Image Detector for the World-Wide Web", *SPIE Vol. 3527*, pp. 68-67, 1998.

[2] D. Doermann, H. Li, and O. Kia, "The Detection of Duplicates in Document Image Databases," *4th Intl. Conf. on Document Analysis and Recognition*, 1997.

[3] A. Hampapur, and R. Bolle "Comparison of Distance Measures for Video Copy Detection," *IEEE ICME 2001*.

[4] A. Jaimes, S.-F. Chang, and A. C. Loui, "A New Framework for Detecting Non-Identical Duplicate Images," *Columbia University Advent Technical Report No. 004, NY, 2002*.

[5] A. Kuchinsky, et al., "FotoFile: A Consumer Multimedia Organization and Retrieval System", *ACM Conf. On Computer and Human Interaction*, Pittsburg, PA 1999.

[6] R. Leinhart, W. Effelsberg, and R. Jain, "VisualGREP: A Systematic Method to Compare and Retrieve Video Sequences," *SPIE Vol. 3312*, pp. 271-282, Jan. 1998.

[7] A.C. Loui and A.E. Savakis, "Automatic Image Event Segmentation and Quality Screening for Albuming Applications," *ICME 2000*, Vol. 2, pp. 1125-1128, 2000.

[8] F. Schaffalitzky and A. Zisserman, "Multi-view Matching for Unordered Image Sets, or "How Do I Organize My Holiday Snaps?"," in proc. *ECCV 2002,* Vol. 1, 2002.

[9] J. Platt, "Auto Album: Clustering Digital Photographs Using Probabilistic Model Merging," in Proc. *IEEE Workshop on Content-Based Access of Image and Video Libraries*, 2000.

[10] L. Wenyin, Y. Sun, H. Zhang, "MiAlbum-A System for Home Photo Management Using the Semi-Automatic Image Annotation Approach," in proc. *ACM Multimedia 2000.*