CAMERA NOTES

Xian-Sheng HUA, Shipeng LI, Hong-Jiang ZHANG
Microsoft Research Asia
{xshua; spli; hjzhang}@microsoft.com

ABSTRACT

Taking notes is frequently required in daily life. The rapid development of consumer devices provides new ways to achieve this goal such as taking notes by digital cameras, camcorders, camera phones, or voice recorders. In this paper, a novel camera-based note taking system is presented, which enables efficient and nature notes taking, management, searching and exporting. Firstly the system automatically classifies photos and video clips taken by a variety of capturing devices into two classes, notes and non-notes, and then further classifies the class of notes into more fractional classes such as document, map, slides, bulletin board, and so on. Next the visual quality of the note photos and video clips are enhanced and adjusted by a set of image and video processing algorithms, as well as textual, color and texture information are extracted from them. And last, based on these analyses, a management system enabling efficient note importing, searching, browsing and exporting is built.

1. INTRODUCTION

Rather than using pen, rapid growth of electronic devices enables users to take notes by a variety of methods including typing, digital ink, cameras, audio recorders, and so on. Among these, camera and audio based notes taking approaches are becoming more and more frequently used due to the convenience of recording and storing. However, it is observed that few of the users will review these notes thereafter. The key issue behind this is the difficulty of information retrieval from their media collections, as well as the visual or aural quality of these electronic notes is typically not sufficiently good for further application or even for reading or listening.

This paper addresses this issue based on multimedia content analysis and processing. In particular, this paper proposes a camera notes collecting, classification, management, browsing and searching system, while aural notes are not considered here. Figure 1 illustrates the flow chart of the proposed system, named *Camera Notes*. Camera Notes consists of three primary components, Notes Classification, Quality Enhancement and Adjustment, and Notes Management.

A closely related research topic to Camera Notes is video and image semantic classification. However, most of previous works are interested in classifying semantic concepts such as indoor-outdoor, cityscape-landscape, or well-defined events in sports, news videos or movies [6][7][10][11], instead of classifying photos and video clips captured for the purpose of taking notes. Video and photo management systems [4][9] are also closely related to Camera Notes. However, we are more focused on *note* photos and video clips. Another related work is lecture or meeting video analysis, in which lecture video segments are aligned with the original slides automatically by content matching [8], or the meeting videos are graphically represented based on memory cues [2]. Text detection in scene image or video is also related to Camera Notes, while our goal is much more than text detection and recognition. We aims at providing a comprehensive notes management system enabling efficient camera, searching, browsing and exporting.

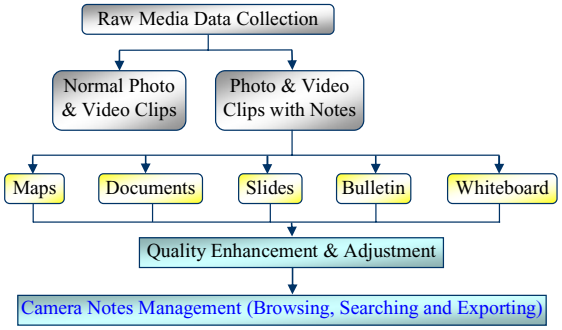


Figure 1. Flow chart of Camera Notes system.

The rest of the paper is organized as follows. Section 2 describes the methods for notes classification, in which photos and video clips regarding notes are separated from normal media collection and then classified into different classes according to their contents. Section 3 presents quality improvement and adjustment for the note photos and video clips, followed by introducing the entire Camera Notes system in Section 4. Section 5 and 6 presents the experimental results and conclusion remarks, respectively.

2. NOTES CLASSIFICATION AND GROUPING

For convenience, photos and video clips that are captured for the purpose of taking notes are called *camera notes*, or *note photos*/*clips*, or simply call them *notes* if there is no confusion in context.

2.1 Note/Non-Note Classification

Note photos and clips taking by cameras (DCs, DVs or phone/PDA cameras) are generally mixed up with normal pictures or video segments. It is labor-intensive and time-consuming for users to manually pick out these note photos and clips. Therefore, the first step is to automatically separate notes from the normal media collection. It should be noted here that the basic classification unit here is one photo or one video shot. Therefore, shot detection based on color similarity or timestamp [9] is applied before doing further analysis and processing.

0-7803-9332-5/05/$20.00 ©2005 IEEE

It is observed that note photos and clips have some distinguish features in comparison with normal ones. Generally their edge distribution is more uniform, as typically they have texts and/or graphs. And, typically the background of notes are (locally) monotony or near monotony so the text or graph could be sufficiently clear. Furthermore, the color distribution of notes photos/clips is often far from balanced, which means, frequently there are only a few dominant colors in them. And, for videos, notes clips are often with mild or even no camera motions. Based on these observations, to separate notes from normal media collection, a set of low-level features are extracted as below.

(1) *Edge Intensity Distribution* (*EID*) vector: The image ($W\times H$) I (photo or video frame) is equally divided into $N\times N$ blocks (denoted by B_i, $0 \le i < N^2$). Then Canny edge detector is applied on the image to form edge map of I, denoted by $E(I)$. The number of edge pixels in block B_i is denoted by $|E(B_i)|$, and let $|E(I)|=\Sigma| E(B_i)|$. The first N^2 elements of *EID* vector is $\{|E(B_i)| / |E(I)|, 0 \le i < N^2\}$. The overall edge ratio over the whole image, i.e., $|E(I)| / (W\times H)$, is taken as the last element of EID vector, thus EID is a (N^2+1)-dimensional vector. For a video shot, the average *EID* over all frames in the shot is taken as the *EID* vector of the shot.

(2) *Background Intensity Distribution* (*BID*) vector: Edge map $E(I)$ is dilated so the text strokes or curves are mostly marked is the dilated map. The complement of the dilated edge map is denoted by $B(I)$, which mostly consists of background pixels. After converting it into grayscale, similar to (1), it is divided into blocks and from which (N^2+1)-dimensional *BID* vector is obtained.

(3) *Color Distribution* (*CD*) vector: 256-dimensional quantized histogram in HSV color space is extracted as *CD* vector.

(4) *Camera Motion Speed* (*CMS*): For video clips, the speed of the camera motion is extracted using the method in [3]. For photos, it is set to zero.

A SVM-based classifier is constructed on a set of pre-labeled samples, and then applied to classify other samples. The performance of this classifier will be presented in Section 5.

2.2 Type Classification

To enable efficient browsing and obtain better results in the process of quality enhancement, another step of classification, note type classification is required. In this step, the camera notes are classified into five typical types including map, document, slides, bulletin and whiteboard, based on the following set of features, except for the feature introduced in Section 2.1:

(1) *Edge Orientation Distribution* (*EOD*) vector: In edge map $E(I)$, the normalized edge orientation histogram (equally quantized into 36 bins) is taken as *EOD* vector.

(2) *Text Ratio, Text Height and Height Variance* (*RHV*): Text detection algorithm proposed in [1] is applied, in which a series text blocks will be obtained. From these blocks, text ratio (the total area of all text blocks over the area of the image), text height (the average height of the blocks), and the variance of the text height are obtained. Therefore, *RHV* is a 3-dimensional feature.

Though the features we used in Section 2.1 are originally for note/non-note classification, it is observed they also facilitate distinguishing different types of notes. Similar to the method for separating note photo/clips from normal media data, five one-versus-rest SVM classifiers are constructed to divide the notes into the five classes mentioned above (according to the largest positive distance from the five classifiers). Experimental results will be introduced in Section 5.

2.3 Grouping and Linking to Normal Media Data

Another issue about camera notes is note grouping. Typically users may take a series of photos or video clips for the same item that needs to be recorded as a note. And, sometimes users may also take some normal photos about the same item when taking notes. Note grouping will group relative note photos and/or clips into groups, as well as generate links from note groups to related photos/video clips in normal media data set.

In Camera Notes, note photos and video clips that are sufficiently close in terms of color similarity and/or timestamps (the time when taking the photos/clips) are regarded as in the same group, and those in normal media collection will be linked to corresponding note group. Users may adjust the grouping and linking results manually through the notes management interface, which will be detailed in Section 4. Figure 2 shows several typical note samples of various types.

Figure 2. Typical note photos (*Top Row*: Map, Document, and Slides; *Bottom Row*: Bulletin, Bulletin and Whiteboard).

3. QUALITY ENHANCEMENT AND ADJUSTMENT

A large number of camera notes are in low quality, whether due to the low resolution of the capture devices, the poor environmental lighting condition, perspective distortion, or nonprofessional shooting skill. Therefore, they are difficult to be directly applied in further applications. In this section, a set of quality enhancement or adjustment algorithms are presented. As most of these methods are typical video/image processing algorithms, they are only briefly introduced here.

Generally full automatic correction methods cannot work well enough for considerable amount of camera notes. Therefore, in our system, we support both automatic methods and semi-automatic ones. The automatic algorithms may be applied automatically or when requested by users. Semi-automatic methods can be applied when necessary manual inputs from users are provided.

3.1 Color/Brightness/Contrast Correction

Manual correction for color, brightness, and contrast, which is similar to those in typical image processing software such as Microsoft Office Picture Manager, PhotoShop, etc., is adopted in the system. Automatic or semi-automatic methods for these corrections have not been integrated into Camera Notes currently, which will be one of our future work items.

3.2 Shape Correction

In [5], a scheme enabling automatically detection and then adjustment for the shape of quadrilateral signboard using perspective transformation is proposed. In Camera Notes, similar technology is adopted. However, the informative areas in many camera notes cannot be automatically detected and corrected. Therefore, a semi-automatic method is provided, in which four vertices of a rectangle in the note photo/video frame are indicated by users, and then a perspective transformation is applied automatically to restore the shape of the rectangle. The users may choose to apply automatic method first, and if failed the semi-automatic method is then used. Figure 3 shows an example note corrected by automatic method.

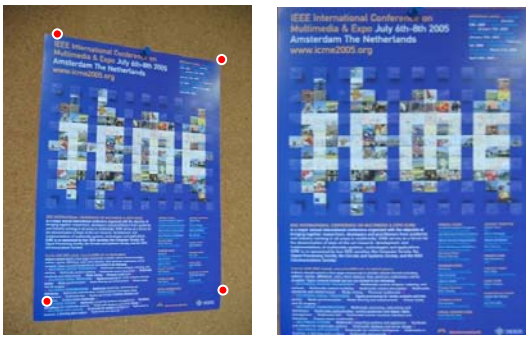


Figure 3. A sample of note shape correction (*Left*: Original. The vertices are detected automatically; *Right*: Corrected).

3.3 Mosaic

In some cases, the item to be captured as note cannot be put into one photo, so multiple photos, or video clips may be applied to capture every part of the map, document, and so on. In these cases, automatic mosaic method is firstly applied to try to recover the original full image. If failed, a semi-automatic method is applied, in which users are required to roughly arrange the position of the photos according to the original, and then a mosaic algorithm is applied.

4. CAMERA NOTES SYSTEM

Based on above analyses and processing, a camera notes management system is built. To provide better search results, text in the notes are detected and recognized using a similar method as the one in [1]. Users may also input keywords manually.

The metadata of notes in the system are described by a XML file, as the example showed in Figure 4. The root element is "CameraNotes", which may consist of one or more sub-elements called "Collection". In "Collection", there are one or more "Group", which contain one ore more "Photo" or "Clip". Except for the root element, for each element in each level, there are several basic properties including *Title*, *STime* (Start Time), *ETime* (End Time), and *Time* (for photos). Element "Photo" and "Clip" include the following sub-elements:

(1) *Src*: URL or file path of the source photo or clip.
(2) *Type*: Type of the note photo or clip, i.e., document, map, and so on.
(3) *Keywords*: The detected and recognized text in the note photo or clip. May also contains users' inputs.
(4) *Links*: Links to related normal photos or clips, which includes one or more photo or video clip URLs, as illustrated in Figure 4.
(5) *ClrHist*: Quatized HSV color histogram of the photo or clip. *ClrHist* is data element, which enables color similarity search in the Camera Notes system.

```
<?xml version="1.0" ?>
<CameraNotes>
    <Collection Title="" STime="" ETime="">
        <Group Title="" STime="" ETime="">
            <Photo Title="" Time="">
                <Src></Src>
                <Type></Type>
                <Keywords></Keywords>
                <![CDATA[ … ]]>
            </Photo>
            <Clip Title="" STime="" ETime="">
                <Src></Src>
                <Type></Type>
                <Keyword></Keyword>
                <![CDATA[ … ]]>
                <Links>
                    <Link Type="Video" Src="" STime="" ETime="">
                    <Link Type="Photo" Src="" Time="">
                </Links>
            </Photo>
        </Group>
    </Collection>
</CameraNotes>
```

Figure 4. Metadata Description.

With these metadata, Camera Notes system enables the following functionalities:

(1) *Note Importing*: The system can be regarded as an additional component or sub-system of MyVideo and MyPhoto [4][9]. It enables automatic (based on note/non-note classification) or manual importing note photos or video clips into note database from raw photo/video collection, or from media database of MyVideo and MyPhoto.
(2) *Sorting and Browsing*: The notes can be sorted or grouped by type, timestamp, and/or key-words. And, users may browse any group of notes in any order. Furthermore, related normal photos and clips can also be easily accessed through the embedded links.
(3) *Editing*: Enables automatic, semi-automatic, and manual quality enhancement/adjustment, mosaicing, etc. Also enables manually group/class classification correction and keywords input.
(4) *Searching*: Provides searching via key-words, note type, time, picture size, and dominant color, or any combination (and/or) of there features.
(5) *Exporting*: Enables automatic creating Word or PDF files, or web pages, as well as printing out an arbitrary selection of the notes.

5. EXPERIMENTS

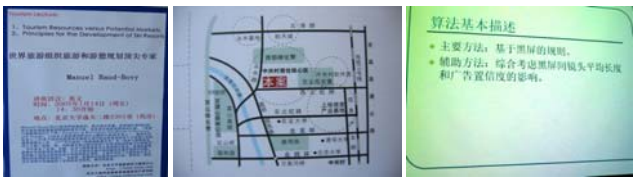
A media database consists of 5 hours of home videos (totally 895 shots) and 1300 photos (830 by typical DCs and 470 by phone cameras) are applied in our experiments, among which there 49

note shots (clips) and 1072 note photos. Table 1 shows the detailed information of the dataset. We randomly choose half of the data as training samples, and other half as test data, and then switch test and train data. Table 1 shows the average precision, recall and accuracy rates of note/non-note classification, and note type classification. From these, it can be seen that note/non-note classification has high accuracy than note type classification. Note type "Slides" has highest recall and precision rates, and the results of whiteboard and map classification are also acceptable. While precision of document and recall of bulletin are quite low. The main reason for this is a number of bulletin notes are similar to document notes, thus they are misclassified. Figure 5 shows several misclassified samples.

Table 1. Precision and Recall of Note Classification.

	Real # (in Total #)	Recall	Precision
Note	1121 (in 2195)	**88.59%**	**91.99%**
Map	176 (in 1121)	84.21%	90.14%
Document	187 (in 1121)	87.50%	48.61%
Slides	383 (in 1121)	98.54%	98.54%
Whiteboard	178 (in 1121)	92.86%	86.67%
Bulletin	197 (in 1121)	50.00%	83.33%
Average (Type)	-	**82.62%**	**81.46%**

Note: "Real # (in Total #)" means the number of real notes, maps, etc. samples ("Real #") in the entire testing set with "Total #" samples. "Average (Type)" indicates averaging results of type classification.



Bulletin/Document *Map/Document* *Slides/Whiteboard*

Figure 5. Misclassified Samples (*Real Type/Classified*).

Figure 6 shows the main UI of the proposed Camera Notes system. This system significantly improves the efficiency of accessing note photos and clips. A rough user study shows that it costs only about 10% of time for locating a specific note in the system by using this system, in comparison with using existing browsing system.

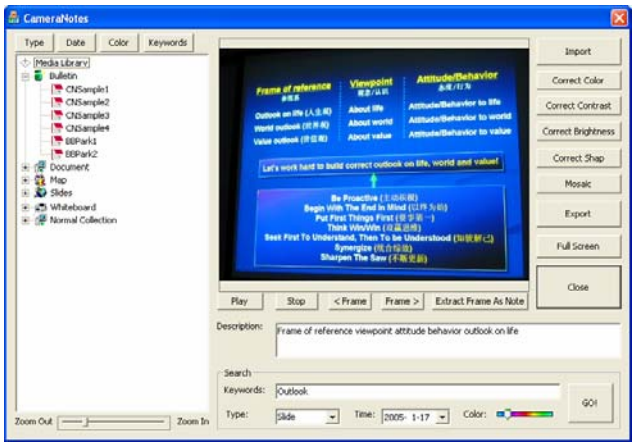


Figure 6. A Snapshot of Camera Notes System.

6. CONCLUSION AND FUTURE WORK

In this paper, Camera Notes, an electronic notes management system based on image and video content analysis is presented. This system enables efficient importing, indexing, browsing, searching and exporting for digital notes taken by typical capturing devices such as camcorders, digital cameras, and camera phones.

Future work includes: (1) To design more and better automatic and semi-automatic quality enhancement/adjustment schemes; (2) To improve the classification accuracy (for both notes/non-notes classification and note type classification) and test on larger dataset; (3) To support more note types; (4) To provide better user interface for the whole camera notes system; (4) To enable more intuitive, convenient and accurate note searching; (5) To support exporting templates enabling flexible and personalized outputs.

7. REFERENCES

[1] X.S. Hua, et al, "Automatic Location of Text in Video Frames," *Workshop on Multimedia Information Retrieval (MIR 2001)*, October 5, Ottawa, Canada. 2001.

[2] A. Jaimes, *et al*, "Memory Cues for Meeting Video Retrieval," *First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, 2004.

[3] D.J. Lan, Y.F. Ma, H.J. Zhang, "A Novel Motion-Based Representation For Video Mining," *IEEE Intl Conf on Multimedia and Expo 2003*.

[4] Y. Sun, H.J. Zhang, L. Zhang, and M. Li. MyPhotos - A system for home photo management and processing. *ACM Multimedia 2002*.

[5] A. Tam, *et al*, "Quadrilateral Signboard Detection and Text Extraction," *Intl Conf on Imaging Science, Systems and Technology 2003*.

[6] A. Vailaya, *et al*, "Content-Based Hierarchial Classification of Vacation Images," *IEEE Intl Conf on Multimedia Computing and Systems*, Jun. 1999.

[7] A. Vailaya, A. Jain, H.J. Zhang, "On Image Classification: City vs. Landscape," *IEEE Workshop on Content - Based Access of Image and Video Libraries*, June 21 - 21, 1998, Santa Barbara, California.

[8] F. Wang, C.W. Ngo, and T.C. Pong, "Synchronization of Lecture Videos and Electronic Slides by Video Text Analysis," *ACM Multimedia 2003*, Berkeley, CA, USA

[9] Y. Wang, P. Zhao, D. Zhang, M. Li, and H.J. Zhang. MyVideos - A system for home video management. *ACM Multimedia 2002*.

[10] Jun WU, *et al*, "An Online-Optimized Incremental Learning Framework for Video Semantic Classification," *12th ACM International Conference on Multimedia*, New York City, USA, Oct. 2004.

[11] Yun Zhai, Zeeshan Rasheed, Mubarak Shah, "A Framework for Semantic Classification of Scenes Using Finite State Machines," *Intl Conf on Image and Video Retrieval*, Dublin, Ireland, July 21-23, 2004.