The Second International Workshop on Mobile Multimedia Processing

In conjunction with The 20th International Conference on Pattern Recognition (ICPR 2010)

Istanbul, Turkey
August 22nd, 2010

Editors

Xiaoyi Jiang,
University of Münster, Germany

Matthew Ma,
Scientific Works, USA

Michael Rohs,
Technical University of Berlin, Germany

ISSN 2191-1517

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

Preface

The portable device and mobile phone market has witnessed rapid growth in the last few years with the emergence of several revolutionary products such as mobile TV, converging iPhone and digital cameras that combine music, phone and video functionalities into one device. The proliferation of this market has further benefited from the competition in software and applications for smart phones such as Google's Android operating system and Apple's iPhone AppStore, stimulating tens of thousands of mobile applications that are made available by individual and enterprise developers. Whereas mobile devices have become ubiquitous in people's daily life not only as a cellular phone but also as a media player, a mobile computing device, and a personal assistant, it is particularly important to timely address challenges in applying advanced pattern recognition, signal, information and multimedia processing techniques, and new emerging networking technologies to such mobile systems.

One attempt on such discussions is the organization of the International Workshop of Mobile Multimedia Processing (WMMP). The primary objective of this workshop series is to foster interdisciplinary discussions and research in mobile multimedia processing techniques, applications and systems, as well as to bring stimulus to researchers on pushing the frontier of emerging new technologies and applications. After the successful first WMMP in Tampa, Florida, in 2008, the second WMMP was held in Istanbul on August 22nd, 2010, in conjunction with the 20th International Conference on Pattern Recognition (ICPR 2010). The proceedings of this workshop contain eleven papers presented at the workshop. The intended readers of the proceedings are primarily researchers wanting to extend traditional information processing technologies to the mobile domain or deploy any new mobile applications that could not be otherwise enabled traditionally.

Our acknowledgment goes to the reviewers who have participated in our stringent reviewing process. We also like to extend our gratitude to all authors for their contributions.

Xiaoyi Jiang
Matthew Ma
Michael Rohs

Proceedings

(in alphabetical order by the first author's name)

A Hybrid Multi-Layered Video Encoding Scheme for Mobile Resource-Constrained Devices

Naveen Kumar Aitha and Suchendra M. Bhandarkar

Department of Computer Science
The University of Georgia
Athens, Georgia 30602-7404, USA
{aitha,suchi}@cs.uga.edu

Abstract. The use of multimedia-enabled mobile devices such as pocket PC's, smart cell phones and PDA's is increasing by the day and at a rapid pace. Networked environments comprising of these multimedia-enabled mobile devices are typically resource constrained in terms of their battery capacity and available bandwidth. Real-time computer vision applications typically entail the analysis, storage, transmission, and rendering of video data, and are hence resource-intensive. Consequently, it is very important to develop a content-aware video encoding scheme that adapts dynamically to and makes efficient use of the available resources. A Hybrid Multi-Layered Video (HMLV) encoding scheme is proposed which comprises of content-aware, multi-layer encoding of the image texture and motion, and a generative sketch-based representation of the object outlines. Each video layer in the proposed scheme is characterized by a distinct resource consumption profile. Experimental results on real video data show that the proposed scheme is effective for computer vision and multimedia applications in resource-constrained mobile network environments.

1 Introduction

The modern era of mobile computing is characterized by the increasing deployment of broadband networks coupled with the simultaneous proliferation of low-cost video capturing and multimedia-enabled mobile devices, such as pocket PC's, smart cell phones and PDA's. Mobile computing has also triggered a new wave of mobile Internet-scale multimedia applications such as video surveillance, video conferencing, video chatting and community-based video sharing, many of which have found their way in practical commercial products. Mobile network environments, however, are typically resource constrained in terms of the available bandwidth and battery capacity on the mobile devices. These environments are also characterized by constantly fluctuating bandwidth and decreasing device battery life as a function of time. Consequently, it is desirable to have a multi-layered (or hierarchical) content-based video encoding scheme where distinct video layers have different resource consumption characteristics and provide information at varying levels of detail [2].

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

Traditional multi-layered video encoding scheme such as the MPEG-4 Fine Grained Scalability profile (MPEG-FGS), are based on the progressive truncation of DCT or wavelet coefficients [3]. There is an inherent trade-off between the bandwidth and power consumption requirements of each layer and the visual quality of the resulting video, i.e., the lower the resource requirements of a video layer, the lower the visual quality of the rendered video [3]. Note that the conventional MPEG-FGS multi-layered encoding is based primarily on the spectral characteristics of low-level pixel data. Consequently, in the face of resource constraints, the quality of the lower layer videos may not be adequate to enable a high-level computer vision or multimedia application. For a multi-layered video encoding technique to enable a high-level computer vision or multimedia application, it is imperative that the video streams corresponding to the lower encoding layers encode enough high-level information to enable the application at hand while simultaneously satisfying the resource constraints imposed by the mobile network environment.

In this paper, a *Hybrid Multi-Layered Video* (HMLV) encoding scheme is proposed which comprises of content-aware, multi-layer encoding of texture and motion and a generative sketch-based representation of the object outlines. Different combinations of the motion-, texture- and sketch-based representations are shown to result in distinct video states, each with a characteristic bandwidth and power consumption profile. The proposed encoding scheme is termed as *hybrid* because its constituent layers exploit texture-, motion- and contour-based information at both the object level and pixel level. The high-level content awareness embedded within the proposed HMLV encoding scheme is shown to enable high-level vision applications more naturally than the traditional multi-layered video encoding schemes based on low-level pixel data.

A common key feature of computer vision and multimedia applications on mobile devices such as smart phones, PDAs and pocket PC's is video playback. Video playback typically results in fast depletion of the available battery power on the mobile device. Various hardware and software optimizations have been proposed to reduce the power consumption during video playback and rendering [2]. Most of the existing work in this area has concentrated on reducing the quality of the video, to compensate for battery power consumption.

More recently, Chattopadhyay and Bhandarkar [1] have proposed a content-based multi-layered video representation scheme for generating different video layers with different power consumption characteristics. The video representation is divided into two components (i) a *Sketch* component, and (ii) a *Texture* component. The Sketch component has two different representations i.e., *Polyline* and *Spline*. The Texture component comprises of three distinct levels (in decreasing order of perceptual quality) denoted by V_{org}, V_{mid} and V_{base}. A combination of any of the three Texture levels and the two Sketch levels are used to generate six distinct levels of video with different resource consumption characteristics.

In this paper, we extend the work in [1] by effectively increasing the number of perceptual layers in the underlying video representation. This allows for a

much finer degree of control on the underlying resource consumption while ensuring optimal perceptual quality of the rendered video for the computer vision or multimedia application on hand. The overall goal is to optimize the performance of the relevant computer vision or multimedia application within the specified resource constraints. In the proposed HMLV scheme, we retain the sketch component of the multi-layered video representation described in [1]. We enhance the texture component described in [1] by using a Gabor Wavelet Transform (GWT)-based representation for the underlying image texture and by including motion layers. The various texture layers are generated uniformly via progressive truncation of the GWT coefficients. The decomposition of the underlying video into motion layers also allows one to order the objects within the field of view based on their approximate depth from the camera.

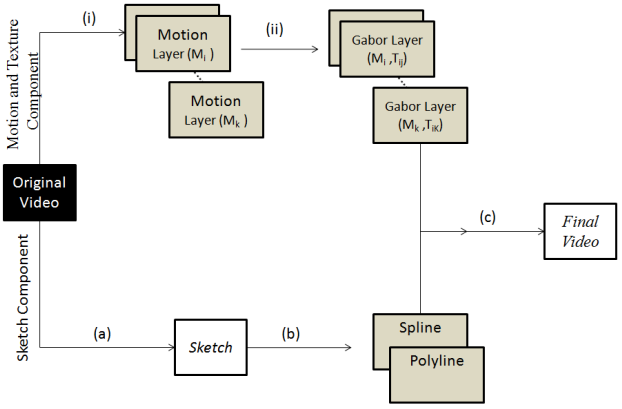
The GWT is a special case of the Short-time Fourier Transform and is used to determine sinusoidal frequency and phase content of local sections of a signal as it changes over time. In recent years, the multichannel GWT has been used for texture analysis and texture representation at multiple scales and orientations. The Gabor filter is a linear filter obtained by computing the GWT coefficients at a specific scale and orientation. A Gabor filter bank is a collection of Gabor filters at multiple scales and orientations. A set of filtered images is obtained by convolving the input image with the bank of Gabor filters. Each of these filtered images represents the input image texture at a certain scale and orientation. The convolution of an input image with a Gabor filter bank bears close resemblance to the processing of images within the primary visual cortex [4].The proposed HMLV scheme is discussed in detail in the following sections.

2 HMLV Encoding

The input video is decomposed into two components: (i) the Sketch component and, (ii) the combined Motion-and-Texture component. The Sketch component is a Generative Sketch-based Video (GSV) representation where the outlines of the objects are represented using sparse parametric curves [1]. The Texture component in [1] is replaced by a combined Motion-and-Texture component in the proposed HMLV scheme since the selection of motion layers is strongly coupled with the process of generating the texture layers via retention (or deletion) of the GWT coefficients for the chosen motion layers. The Texture-and-Motion component in the proposed HMLV scheme is represented by four layers, i.e., a base layer video, two intermediate mid-layer videos and the original video. The combination of the Sketch component and different Motion-and-Texture layer videos (base, mid-level or the original video) yields distinct video states with unique resource utilization characteristics. Figure 1 shows the proposed HMLV scheme.

2.1 Encoding $V_{Motion-and-Texture}$

Multi-scale representation of the image texture is achieved using GWT coefficients at different scales and orientations. The generation of different Motion-



HLV Encoding Scheme: (i) Motion Layers(k) are generated for the Motion and Texture component, (ii) The chosen k motion layers are encoded at different texture level, (a) Sketch component extracted from the video, (b) Polyline and Spline representations are generated, (c) One of the sketch representations is chosen for each of k motion layers superimposed together to form the final video

Fig. 1. Hybrid Multi-Layered Video Encoding Scheme

and-Texture layers is dependent on two factors, i.e., the number of Motion layers selected and the truncation parameter for the GWT coefficients (i.e., Texture level) used to represent each Motion layer.

The Motion-and-Texture component of the HMLV representation comprises of four distinct layers denoted by: V_{base}, V_{I1}, V_{I2} and V_{orig}, where V_{base} is the lowest-level layer encoded using the fewest GWT coefficients for all the Motion layers; V_{orig} is the highest layer which is represented using the maximum number of GWT coefficients for all Motion layers resulting in a video of the highest visual quality; and V_{I1} and V_{I2} are the mid-level layers where the Motion layers that are deemed important are encoded using more GWT coefficients and the rest using fewer GWT coefficients.

2.2 Generation of Motion Layers

The original video is first processed to extract the different Motion layers, which form the basis for the generation of the Motion-and-Texture component. For any two successive video frames, the motion parameters for the Tomasi and Shi feature points [9] are estimated using an optical flow function [10]. After estimating the motion vectors, an adaptive K-means clustering technique [8] is used to cluster the motion vectors. Assuming that image pixels belonging to a single object share similar motion, spatial information is exploited in the clustering of the motion vectors to generate distinct Motion layers. The background motion layer is assumed to have constant or zero motion, and the remaining layers to have non-zero motion. A novel Motion-based Multi-Resolution (MMR) encoding scheme is proposed to encode distinct Motion layers at varying levels of visual quality.

2.3 Motion-based Multi-Resolution (MMR) Encoding Scheme

Each frame is represented by selecting the relevant Motion layers and the GWT coefficient truncation parameter that is used to encode each Motion layer. The GWT coefficient truncation parameter for each Motion layer is chosen from {0.0, 0.5, 0.7, 0.8, 0.9, 1.0}. The Motion layer is then encoded with the corresponding number of GWT coefficients. Let F_i be the Motion-and-Texture frame of the video which is to be combined with the Sketch component to constitute the final video. Let $M_{i1}, M_{i2}....M_{ik}$ be the k Motion layers to be encoded in frame i. Let $T_{i1}, T_{i2}....T_{ik}$ be the k texture levels (based on the GWT truncation parameters) for the k motion layers found in frame i. Then (M_{ij}, T_{ij}) represents the motion layer j of frame i which is encoded using the Texture level T_{ij}. The final Motion-and-Texture frame is generated via overlay of all the Motion layers. i.e.,

$$F_i = \sum_{j=0}^{j=k} (M_{ij}, T_{ij})$$

where $0 \leq T_{ij} \leq 1$ is the normalized value for the GWT coefficient truncation parameter.

2.4 Generating the Highest (Top-most) Layer V_{orig}

Using the above MMR encoding scheme, the original video can be generated using all the GWT coefficients, i.e., $T_{ij} = 1, \forall i, j$. This is tantamount to the encoding of each of the Motion layers with all the GWT coefficients resulting in full reconstruction of each frame in the video stream.

2.5 Generating the Base Layer V_{low}

The lowest-level Motion-and-Texture layer can be generated using very few GWT coefficients for all the motion layers, i.e., $T_{ij} = 0.5, \forall i, j$. Using only a few GWT coefficients results in a smoothed reconstruction of the Motion layer, which, when overlayed with the Sketch component, improves the visual appeal of the frame. Deleting entirely the background Motion layer in V_{low} generates the layer V_{base0} which is deemed to have a lower power consumption profile than the base layer.

2.6 Generating Intermediate Layers V_{mid}

Most of the commonly video available encoding techniques are not content-aware, but in the proposed HMLV encoding scheme the intermediate layers are designed in such a manner that they articulate perfectly the high-level contents of the video and encode the information accordingly. The two intermediate layers are generated as follows:
(i) Encode the motion layer which is farthest from the camera with zero GWT coefficients and all the other motion layers including background motion layer with the maximum number of GWT coefficients. (i.e., $T_i = 1$).
(ii) Encode the motion layer corresponding to the background at a low Texture level (i.e., with the fewest GWT coefficients) and all other layers at a high Texture level (i.e., $T_i = 1$).

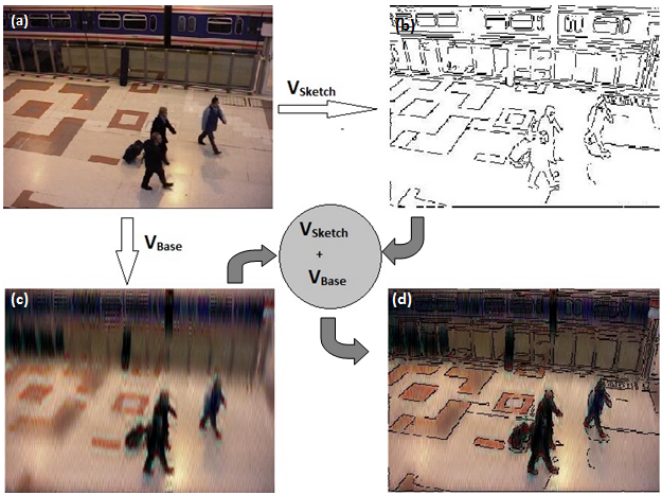


Fig. 2. Example of the Generation of a Hybrid frame from the Base Layer and Polyline Sketch (a) Original Frame, (b) Sketch Frame, (c) Motion-and-Texture Frame, (d) HMLV frame

3 Combining V_{Sketch} and $V_{Motion-and-Texture}$

In the proposed HMLV scheme, $V_{Motion-and-Texture}$ and V_{Sketch} are obtained independently of each other. A suitable Motion-and-Texture frame is generated and written to the frame buffer by the video controller. The Sketch component is extracted subsequently and superimposed on the Motion-and-Texture frame. The components are processed independently; only the order in which they are rendered is different. The Motion-and-Texture frame is rendered first followed by the superimposition of the Sketch frame. The video states resulting from different combinations of the Motion-and-Texture component and the Sketch component are arranged in a linear order starting with the highest quality video and ending with the lowest quality video. The first video state is deemed to consume the most battery power and the last video state the least battery power. The experimental results presented in the next section demonstrate that the distinct video states indeed have different power consumption characteristics.

4 Results

This paper discusses the efficient representation of the motion and texture within a video frame using motion layers and the GWT coefficient truncation parameter resulting in a simple HMLV encoding framework. In Figure 2, we can observe the generation of a hybrid frame from a Base Layer frame and Polyline Sketch frame. Figure 3 shows different Motion-and-Texture layers associated with a single frame that are generated using the proposed HMLV encoding scheme, i.e.,

Fig. 3. Different Motion-and-Texture Levels (a) Base Layer with Background Motion Layer Removed (b) Base Layer with GWT Coefficient Truncation Parameter = 0.5 (c) Intermediate Layer (d) Original Layer with GWT Coefficient Truncation Parameter = 1.0

the base layer, the intermediate layers and the original layer. We can see that the final visual appeal of the frame is increased by the overlay of the Sketch component over the base and intermediate Motion-and-Texture layers. The base layer and the intermediate Motion-and-Texture layers when overlaid with the Sketch component can be effectively used in video surveillance and object tracking applications where the finer details of the video are irrelevant.

The power consumption profiles for different video states are shown in Figure 4. The overall power consumed during the video playback process is calculated in terms of the available battery time. It can be seen from the graph that the lower levels of texture take less battery power than the higher levels. The V_{Sketch} layer followed by V_{Base0} layer (i.e., the base level texture without the background motion layer) consume the minimum amount of battery power whereas the original video consumes the maximum amount of battery power on the device. All experiments have been performed using a Dell Inspiron 1525 laptop PC with 2.0GHZ CPU, 2GB RAM, and a 250GB, 5400 rpm hard drive running in battery mode.

5 Conclusions

This paper presents an integrated HMLV representation framework for contour-, motion- and texture-based encoding using object outline sketches, motion layers and GWT coefficient truncation parameters. The earlier HLV encoding scheme [1] generated only three texture levels V_{orig}, V_{mid} and V_{base} resulting in six distinct HLV encoding levels, whereas the proposed HMLV encoding technique generates four Motion-and-Texture levels - the original layer, two intermediate layers and a base layer. This results in more fine-grained HMLV levels which make more efficient use of the available resources. This is very much evident from the power consumption profiles of different HMLV layers. we plan to evaluate the proposed HMLV scheme in the context of some important multimedia applications in a resource-constrained mobile network environment, such as face detection, face recognition and face tracking, using quantitative metrics.

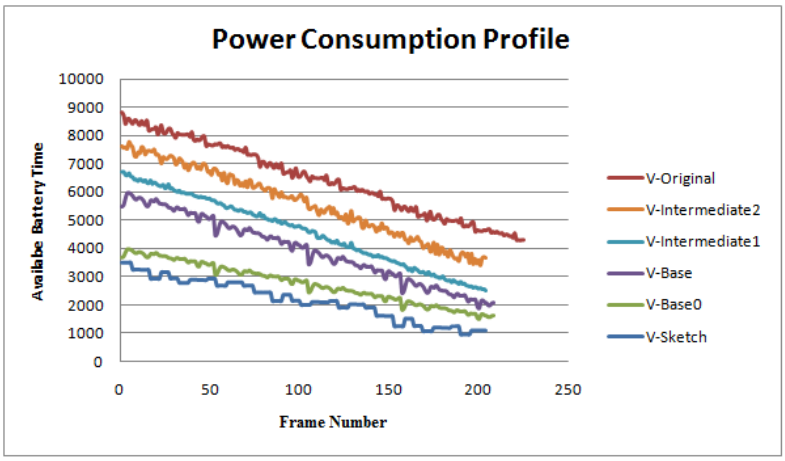


Fig. 4. Power Consumption Profiles for Different Texture Levels

References

1. Chattopadhyay S, Bhandarkar S. M. Hybrid layered video encoding and caching for resource constrained environments, *Jour. Visual Communication and Image Representation*, 19(8):573-588, 2008.
2. Sikora, T. Trends and prespectives in image and video coding, *Proceedings of the IEEE*, 93(1):6-17, 2005.
3. Dai, M., Loguinov, D. Analysis and Modeling of MPEG-4 and H.264 Multi-Layer Video Traffic, *Proceedings of IEEE Infocom*, 2005.
4. Lee, T.S., Image Representation using 2D Gabor Wavelets, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10) 1996.
5. Hakeem, A., Shafique, K., Shah, M. An object-based video coding framework for video sequences obtained from static cameras, *Proceedings of the 13th Annual ACM International Conference on Multimedia*, 608-617, 2005.
6. Chattopadhyay S, Bhandarkar S. M, Li K. FMOE-MR: content-driven multi-resolution MPEG-4 fine-grained scalable layered video encoding, *Proc. ACM Multimedia Computing and Networking Conference*, 650404-1- 11, 2007.
7. Ku, C.-W., Chen, L.-G., Chiu, Y.-M. A Very Low Bit-Rate Video Coding System based on Optical Flow and Region Segmentation Algorithms, *Proceeding of the SPIE Conf. Visual Communication and Image Processing, Taipei*, 3:13181327, 1995.
8. Chen, H., Wu, X., Hu, J. Adaptive K-Means clustering Algorithm, *Proc. of SPIE*, 6788,67882A(2):167-192, 2007.
9. Shi, J., Tomasi, C. Good Features to Track, *IEEE Conference on Computer Vision and Pattern Recognition*, 593-600, 1994.
10. Lucas, B. D., Kanade, T. An Iterative Image Registration Technique with an Application to Stereo Vision, *International Joint Conference on Artificial Intelligence*, 674-679, 1981.

Survey of SIFT Compression Schemes

Vijay Chandrasekhar*, Mina Makar*, Gabriel Takacs*, David Chen*,
Sam S. Tsai*, Ngai-Man Cheung*, Radek Grzeszczuk†,
Yuriy Reznik‡, and Bernd Girod*

*Stanford University, † Nokia Research Center, CA ‡ Qualcomm Inc., CA

Abstract. Transmission and storage of local feature descriptors are of critical importance for mobile visual search applications. We perform a comprehensive survey of Scale Invariant Feature Transform (SIFT) compression schemes proposed in the literature and evaluate them in a common framework. Further, we compare the different schemes to the recently proposed low bit-rate Compressed Histogram of Gradients (CHoG) descriptor. We show that CHoG outperforms all SIFT compression schemes. We implement CHoG in a large-scale mobile image retrieval system and show that transmitting CHoG feature data are an order of magnitude smaller than transmitting SIFT descriptors or JPEG images.

1 Introduction

Local image features have become pervasive in the areas of computer vision and image retrieval. Feature compression is vital for reduction in storage, latency and transmission in mobile visual search applications.

Transmission time: For mobile visual search applications, bandwidth is a limiting factor. One approach used in mobile visual search applications is to transmit the JPEG compressed query image over the network, but this might be prohibitively expensive at low uplink speeds. An alternate approach is to extract feature descriptors on the phone, compress the descriptors and transmit them over the network. In [1], we show that such an approach can reduce the application latency by an order of magnitude.

Server-client Caching: For several applications, a subset of descriptors are stored in RAM for fast access. Takacs *et al.* [2] cache a set of descriptors on the mobile client to enable an outdoors mobile augmented reality experience. Having compact descriptors allows storage of a larger set of descriptors in main memory.

Server-side Storage: Image and video retrieval applications need query images to be matched against databases of billions of features. E.g., 1000 hours of video produces ∼10 billion local descriptors. Storing 10 billion uncompressed SIFT [3] descriptors would require ∼10 TB of storage. More compact descriptors will lead to faster file accesses.

SIFT is the most popular descriptor in computer vision for retrieval applications. In this work, we perform a comprehensive survey of SIFT compression schemes and evaluate them in a common framework.

1.1 Prior Work and Outline

We broadly classify SIFT compression schemes into three categories: Hashing, Transform Coding, and Vector Quantization.

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

Hashing: Locality Sensitive Hashing (LSH) [4, 5] is the most popular hashing technique for high dimensional descriptors. In [6], Torralba *et al.* build binary codes for high dimensional descriptors using machine learning techniques like Restricted Boltzmann Machines (RBM) and Similarity Sensitive Coding (SSC). In [7], Weiss *et al.* propose a scheme called Spectral Hashing (SH) that outperforms RBM and SSC based approaches. For each of these schemes, exact Euclidean distances are approximated or estimated by Hamming distances over binary codewords.

Transform Coding: In [8], we have studied dimensionality reduction of SIFT descriptors using the Karhunen-Lòeve Transform (KLT) followed by entropy coding. The KLT-based coding is known to work best for data with Gaussian statistics. Since SIFT statistics are highly non-Gaussian, we also study the performance of a transform coding scheme based on Independent Component Analysis (ICA).

Vector Quantization: Vector Quantization (VQ) of SIFT features is most commonly used in the "bag-of-features" image retrieval framework. For such applications, SIFT features are quantized using flat k-means (FKM) or hierarchical k-means (HKM) [9] to form a bag of "visual words". HKM or FKM can also be used for compression of descriptors for storage or transmission. Jegou *et al.* [10] in their recent work propose a scheme called Product Quantization (PQ), where the SIFT descriptor is divided into smaller blocks and VQ is performed on each block. Some hybrid schemes have also been proposed in the literature. In [11], Jegou *et al.* propose a scheme called Hamming Embedding (HE), where SIFT descriptors are first coarsely quantized using HKM, and binary hashes are used for refinement in each quantization cell.

In our own work [12, 1], we propose a framework for computing low bit-rate feature descriptors called CHoG. Gradient histograms are quantized using Huffman trees, Type Quantization or Lloyd Max VQ and compressed efficiently using fixed or variable length codes.

Here, we will evaluate the different SIFT compression schemes in a common framework. In Section 2, we survey the different SIFT compression schemes proposed in the literature. In Section 3, we review the CHoG descriptor. Finally, in Section 4, we evaluate the performance of the different schemes.

2 SIFT Compression Schemes

2.1 Hashing

Locality Sensitive Hashing: We use the LSH scheme proposed by Yeo *et al.* [4]. To build a hash, we first randomly generate a set of hyperplanes that pass through the origin. Each bit of the hash is then determined by which side of the hyperplane the SIFT descriptor lies. and the Hamming distance of their hashes.

Similarity Sensitive Coding: Torralba *et al.* [6] use machine learning techniques such as SSC and RBM to train binary codes for high dimensional descriptors. The Boosting SSC algorithm learns an embedding of the original Euclidean space into a binary space such that distances between vectors in the original

space are correlated with their Hamming distances in the binary space. Each bit of the hash is obtained as the output of a weak classifier.

Spectral Hashing: Weiss *et al.* propose Spectral Hashing in their recent work [7]. The spectral hashing scheme performs Principal Component Analysis (PCA) on the data and fits a multidimensional rectangle to it. The dimensions of the rectangle determine the hashing functions of each bit.

2.2 Transform Coding

Karhunen-Loève Transform: Transform coding of SIFT descriptors was first proposed in [8]. The compression pipeline first applies a Karhunen-Lòeve Transform (KLT) transform (or PCA) to decorrelate the different dimensions of the feature descriptor. Then, each dimension of the KLT vector is scalar quantized. The quantized coefficients of the descriptors are entropy coded with an arithmetic coder. In [8], it was observed that applying the KLT is effective at low rates, but hurts performance at high rates. At high rates, scalar quantization and entropy coding with no transform performs better.

The KLT gives a good rotation for scalar quantization for Gaussian statistics as the decorrelation aligns the Gaussian distribution with the symbol axes. The KLT is optimal for Gaussian data, causing the transformed coefficients to be statistically independent. However, the statistics for SIFT features are not Gaussian as shown in Figure 1. Our goal is to make the transformed coefficients as statistically independent as possible. Hence, we explore an ICA based transform which outperforms the KLT.

ICA based Transform: Under high-rate assumptions, Narozny *et al.* in [13] provide a framework to compute the optimal linear transform for making the transformed coefficients as independent as possible, without assuming Gaussian statistics of the input signal or orthogonality of the transform matrix. They define the Generalized Coding Gain (GCG) as the ratio between distortion-rate functions in case of no transform vs. applying the transform. The transform that maximizes the GCG also maximizes the Generalized Maximum Reducible Bits defined as

$$\begin{aligned} R_{\text{GMRB}} &= R_{\mathbf{I}}(D) - R_{\mathbf{A}}(D) \\ &= \tfrac{1}{d} I(X_1;\ldots;X_d) - \tfrac{1}{d} I(Y_1;\ldots;Y_d) - \tfrac{1}{2d}\log_2\left(\frac{\det\left[\operatorname{diag}\left(\mathbf{A}^{-\mathrm{T}}\mathbf{A}^{-1}\right)\right]}{\det[\mathbf{A}^{-\mathrm{T}}\mathbf{A}^{-1}]}\right), \end{aligned}$$

where $R_{\mathbf{I}}(D)$ and $R_{\mathbf{A}}(D)$ are rate-distortion functions in case of no transform and applying the transform respectively, d is the length of the descriptor, $\mathbf{X} = [X_1;\ldots;X_d]$ is a vector which represents the signal to be encoded, $\mathbf{Y} = [Y_1;\ldots;Y_d]$ represents the transform coefficients, $I(.)$ is mutual information function and $\mathbf{A}$ is the transform $d \times d$ matrix where $\mathbf{Y} = \mathbf{AX}$. Thus, the optimal linear transform $\mathbf{A_{opt}}$ is calculated as

$$\mathbf{A_{opt}} = \arg\min_{\mathbf{A}} I(Y_1;\ldots;Y_d) + \tfrac{1}{2}\log_2\left(\frac{\det\left[\operatorname{diag}\left(\mathbf{A}^{-\mathrm{T}}\mathbf{A}^{-1}\right)\right]}{\det[\mathbf{A}^{-\mathrm{T}}\mathbf{A}^{-1}]}\right)$$

The first term is non-negative and equals to zero if and only if the transform coefficients are independent. The matrix that minimizes this term is the solution to the ICA problem. Note that the second term is also non-negative and equal

to zero iff the columns of $\mathbf{A}^{-1}$ are pairwise orthogonal. Hence, it is considered as a pseudo-distance to orthogonality of the transform matrix $\mathbf{A}$. We use the code provided by the authors in [13] to solve the minimization problem. For more details, the reader is referred to [13]. The transform is applied the same way as KLT, but we expect better compression efficiency due to the highly non-Gaussian statistics of the SIFT descriptors.

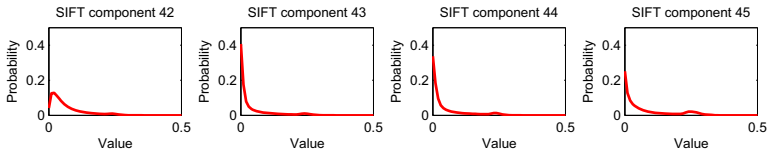


Fig. 1. We observe that the statistics of SIFT descriptors are non-Gaussian.

2.3 Vector Quantization

Product Quantization: Since the SIFT descriptor is high dimensional, Jegou *et al.* [10] propose a product quantizer which operates on lower dimensional subspaces and quantizes each subspace separately. The authors propose two variants of the scheme. The first scheme decomposes the SIFT descriptor directly into smaller blocks and performs VQ on each block. In the second scheme, the authors coarsely quantize the full descriptor with flat k-means or hierarchical k-means using 10^3 to 10^6 nodes. The residual is then quantized using a product quantizer. The two schemes perform comparably, and we consider the former scheme in our comparisons here.

In [10], the authors also investigate how different dimensions of the descriptor should be grouped together for good performance. The order that corresponds to grouping consecutive components together performs the best. Intuitively, this works well because histograms of consecutive cells are quantized together. There are two parameters used to control the bitrate: the number of blocks, denoted as B, and the size of the codebook for each block, denoted as C. Fixed length codes are used for each block and the bitrate of each descriptor is given by $B \times \lceil \log_2 C \rceil$. In Section 4, we consider $B = 1, 2, 4, 8, 16$ and $C = 16, 64, 256, 1024$. In the case of $B = 1$, the product quantizer reverts to flat k-means for the full descriptor.

Tree Structured Vector Quantization: Nistér and Stewénius [9] use HKM or a Tree Structured Vector Quantizer (TSVQ) to quantize SIFT descriptors and build a Inverted File System for fast indexing. Here, we use the same scheme for quantization and compression of SIFT descriptors. We quantize SIFT descriptors with a 10^6 node TSVQ with a branch factor (BF) of 10 and depth (D) of 6, requiring 20 bits per descriptor. A significantly larger TSVQ is not practical due to the size of the code book.

3 CHoG Descriptor

CHoG [12] is a Histogram of Gradients descriptor that is designed to work well at low bitrates. We highlight some key aspects of the descriptor here and readers are referred to [12, 1] for more details. First, we divide the patch into soft log polar spatial bins using DAISY configurations proposed in [14]. Next, the joint

(d_x, d_y) gradient histogram in each spatial bin is captured directly into the descriptor. CHoG histogram binning exploits the skew in gradient statistics that are observed for patches extracted around keypoints. Finally, CHoG retains the information in each spatial bin as a distribution. This allows the use of more effective distance measures like KL divergence, and more importantly, enables efficient quantization and compression. Typically, 9 to 13 spatial bins and 3 to 9 gradient bins are chosen resulting in 27 to 117 dimensional descriptors.

For compressing the descriptor, we quantize the gradient histogram in each cell individually and map it to an index. The indices are encoded with fixed length or entropy codes, and the bitstream is concatenated together to form the final descriptor. Fixed-length encoding provides the benefit of compressed domain matching at the cost of a small performance hit. In prior work [12, 1], we have explored several schemes that work well for histogam compression: Huffman Coding, Type Coding and optimal Lloyd-Max VQ. Here, we use Type Coding, which is linear in complexity to the number of histogram bins and performs close to optimal Lloyd-Max VQ. Readers are referred to [1] for details of the histogram quantization and compression schemes.

4 Results

In Section 4.1, we evaluate the different compression schemes at the feature level using Receiver Operating Characteristic (ROC) curves. In Section 4.2, we compare transmitting CHoG descriptors to SIFT descriptors or JPEG images over a 3G network in a real-world mobile visual application.

4.1 Feature Level Performance

For evaluating the performance of low bitrate descriptors, we use the two data sets provided by Winder and Brown in their most recent work [14], *Notre Dame* and *Liberty*. For algorithms that require training, we use the *Notre Dame* data set, while we perform our testing on the *Liberty* set. From the distances between matching and non-matching pairs of descriptors, we obtain a Receiver Operating Characteristic (ROC) curve which plots correct match fraction against incorrect match fraction. For a fair comparison at the same bitrate, we consider the Equal Error Rate (EER) point on the different ROC curves for each scheme.

Figure 2 and Table 1 summarize the bitrate EER trade-off for different schemes. The number of bits required to match the performance of 1024-bit SIFT is presented in Table 1. For Table 1, note that complexity refers to the number of operations required for compressing each descriptor.

First, we compare the 3 hashing schemes. We note that LSH requires about 1000 bits to match the performance of SIFT, which is close to the size of the uncompressed descriptor itself. SSC and Spectral Hashing perform better than LSH at low bitrates but suffer due to overtraining at high bitrates. At high rates, there is a significant gap in performance between the uncompressed 1024-bit SIFT descriptor and hashing schemes based on machine learning. An advantage of hashing schemes is that they can be compared in the compressed domain using look-up tables.

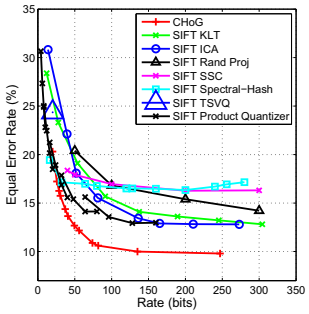


Fig. 2. Comparison of EER versus bitrate for different SIFT compression schemes for the *Liberty* data set. We observe that CHoG outperforms all other schemes.

For transform coding, we observe that the KLT scheme matches the performance of SIFT at about 200 bits. The ICA transform scheme gives a 10-25% reduction in bitrate at a fixed EER compared to the KLT scheme. The transform coding schemes outperform hashing schemes by a significant margin.

The TSVQ compression at 20 bits performs poorly and does not come close to the performance of SIFT. The PQ scheme performs best at low rates. The PQ scheme requires about 160 bits to match the peformance of SIFT as also observed by the authors in [10]. Both ICA and PQ schemes require a bitrate of 160 bits to match the performance of SIFT. Note, however, for PQ at this bitrate, the size of the codebook C=1024, and the scheme is an order of magnitude more complex than transform coding.

Finally, we observe that CHoG outperforms all SIFT compression schemes. Note from Table 1 that CHoG provides several key advantages: no training, significantly lower complexity $O(d)$, and compressed domain matching. The gain in performance comes from using a more compact spatial footprint, KL distance for comparisons and a highly efficient quantization and compression scheme. We conclude that we can achieve better performance with CHoG, which is designed taking compression into account, compared to compressing SIFT.

Scheme	# of bits	Training	Complexity	CDM
LSH	1000	×	$O(Nd)$	√
SSC	-	√	$O(Nd)$	√
S-Hash	-	√	$O(Nd)$	√
KLT	200	√	$O(d^2)$	×
ICA	160	√	$O(d^2)$	×
PQ	160	√	$O(Cd)$	√
TSVQ	-	√	$O(BDd)$	√
CHoG	60	×	$O(d)$	√

Table 1. Results for different compression schemes. CDM is Compressed Domain Matching. N is the number of hash-bits. $d = 128$ for SIFT schemes. C = size of codebook for PQ scheme. B = breadth of TSVQ. D = depth of TSVQ.

4.2 Image Retrieval Performance

We evaluate the performance of CHoG in a large scale mobile image retrieval system. We use a database of one million CD, DVD and book cover images, and a set of 1000 query images [15] exhibiting challenging photometric and geometric distortions. The server retrieval pipeline is based on techniques proposed in [9]. More details can be obtained in [1]. Each image has 500×500 pixels resolution. We define Classification Accuracy (CA) as the percentage of query images correctly retrieved. The data transmission experiments are conducted in a AT&T 3G wireless network, averaged over several days, with a total of more than 5000 transmissions at indoor locations where a image-based retrieval system would be typically used.

Figure 3 compares schemes based on CHoG, SIFT and JPEG. For the JPEG scheme, the bitrate is varied by changing the quality of compression. For SIFT, we transmit uncompressed 1024-bit descriptors, and for CHoG, we transmit 60-bit descriptors. For SIFT and CHoG, we sweep the CA-bitrate curve by varying the number of descriptors transmitted. In Figure 3(*left*), we observe that the amount of data for CHoG descriptors are an order of magnitude smaller than JPEG images or SIFT descriptors, to achieve the same CA. In Figure 3(*right*), we study the average end-to-end latency at the highest accuracy point for the different schemes. We achieve approximately 2-4× reduction in system latency with CHoG descriptors compared to JPEG images or uncompressed SIFT descriptors.

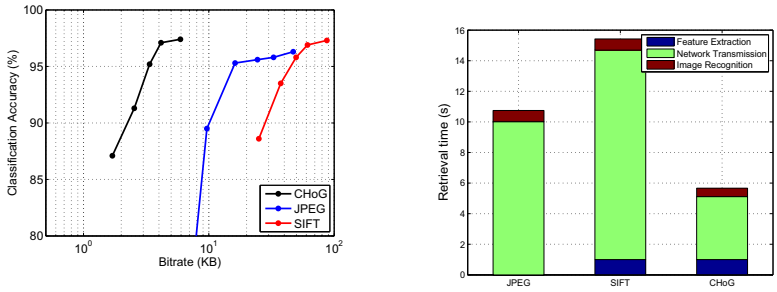


Fig. 3. Figure(*left*) compares birate-classification accuracy of different schemes. CHoG descriptor data are an order of magnitude smaller compared to transmitting JPEG images or uncompressed SIFT descriptors, at the same CA. Figure(*right*) compares end-to-end latency for different schemes. Compared to SIFT and JPEG schemes, CHoG achieves approximately 2-4× reduction in average system latency in a 3G network.

5 Conclusion

We perform a comprehensive survey of SIFT compression schemes and evaluate them in a common framework. We achieve better performance with CHoG, which is designed taking compression into account, compared to compressing SIFT. The CHoG descriptor at 60 bits matches the performance of the uncompressed 1024-bit SIFT descriptor. We evaluate the performance of CHoG in a large-scale mobile image retrieval system, and show that we can achieve 2-4× reduction

in latency by transmitting CHoG descriptors compared to SIFT descriptors or JPEG compressed images.

References

1. Chandrasekhar, V., Reznik, Y., Takacs, G., Chen, D.M., Tsai, S.S., Grzeszczuk, R., Girod, B.: Study of Quantization Schemes for Low Bitrate CHoG descriptors. In: Proceedings of IEEE International Workshop on Mobile Vision (IWMV), San Francisco, California (June 2010)
2. Takacs, G., Chandrasekhar, V., Gelfand, N., Xiong, Y., Chen, W., Bismpigiannis, T., Grzeszczuk, R., Pulli, K., Girod, B.: Outdoors augmented reality on mobile phone using loxel-based visual feature organization. In: Proc. of ACM International Conference on Multimedia Information Retrieval (ACM MIR), Vancouver, Canada (October 2008)
3. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
4. Yeo, C., Ahammad, P., Ramchandran, K.: Rate-efficient visual correspondences using random projections. In: Proc. of IEEE International Conference on Image Processing (ICIP), San Diego, California (October 2008)
5. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Commun. ACM **51**(1) (2008) 117–122
6. Torralba, A., Fergus, R., Weiss, Y.: Small Codes and Large Image Databases for Recognition. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2008)
7. Weiss, Y., Torralba, A., Fergus, R.: Spectral Hashing. In: Proceedings of Neural Information Processing Systems (NIPS), Vancouver, BC, Canada (December 2008)
8. Chandrasekhar, V., Takacs, G., Chen, D.M., Tsai, S.S., Girod, B.: Transform coding of feature descriptors. In: Proc. of Visual Communications and Image Processing Conference (VCIP), San Jose, California (January 2009)
9. Nistér, D., Stewénius, H.: Scalable recognition with a vocabulary tree. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New York, USA (June 2006)
10. Jegou, H., Douze, M., Schmid, C.: Product Quantization for Nearest Neighbor Search. Accepted to IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)
11. Jegou, H., Douze, M., Schmid, C.: Hamming embedding and weak geometric consistency for large scale image search. In: Proc. of European Conference on Computer Vision (ECCV), Berlin, Heidelberg (2008) 304–317
12. Chandrasekhar, V., Takacs, G., Chen, D.M., Tsai, S.S., Grzeszczuk, R., Girod, B.: CHoG: Compressed Histogram of Gradients - A low bit rate feature descriptor. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, Florida (June 2009)
13. Narozny, M., Barret, M., Pham, D.T.: Ica based algorithms for computing optimal 1-d linear block transforms in variable high-rate source coding. Signal Process. **88**(2) (2008) 268–283
14. Winder, S., Hua, G., Brown, M.: Picking the best daisy. In: Proc. of Computer Vision and Pattern Recognition (CVPR), Miami, Florida (June 2009)
15. Chen, D.M., Tsai, S.S., Vedantham, R., Grzeszczuk, R., Girod, B.: CD Cover Database - Query Images. (April 2008)

Mobile OCR on the iPhone for Different Types of Text Documents

Ralph Ewerth, Jan Peer Harries, and Bernd Freisleben

Dept. of Mathematics and Computer Science, University of Marburg,
Hans Meerwein Str. 3, D-35032 Marburg, Germany
{ewerth, harries, freisleb}@informatik.uni-marburg.de

Abstract. Several approaches for solving the problem of optical character recognition (OCR) for machine-printed documents have been proposed in the literature. Typically, these approaches perform well when documents are scanned under controlled illumination and at a resolution of at least 300 dpi, but when documents are captured by digital cameras at lower resolutions and in arbitrary environments, their performance degrades significantly. In this paper, we present an approach for mobile OCR running entirely on Apple's iPhone. The approach is based on the free OCR software tesseract. Several image processing and enhancement techniques are proposed to deal with arbitrary text and illumination conditions. In contrast to previous work, all processing steps are executed on the mobile phone. Experimental results are reported for a test set of 46 images, improving the accuracy of the baseline system by more than 12% and outperforming two commercial OCR applications for the iPhone.

Keywords: OCR, iPhone, mobile media, mobile image processing.

1 Introduction

Optical character recognition (OCR) converts text in a digital image into textual form, i.e. from the pixel domain to ASCII code or Unicode. This allows users to save disk storage, to use the recognized text in word processing or other office applications, and to search for text in databases. OCR has been a field of active research for decades, and the issue of recognizing machine printed characters in scanned text documents can be considered as solved since the mid-nineties. Current research focuses on the recognition of handwritten characters (e.g. [1, 2]), recognizing text of languages with a large alphabet (e.g. [17]), or recognizing text in images and videos (e.g. [5, 6]). For machine printed text of Latin characters, commercial OCR systems are available with low error rates for common text documents.

However, if a digital camera has captured an image containing text, the problem of OCR gets more difficult. First, a captured ISO/DINA4 document should be processed with approximately 300 dots per inch (dpi). For ISO/DINA4, an image resolution of at least 2480*3508 pixels is required, that is 8.7 MegaPixels. Second, in contrast to scanning devices, the illumination is not controlled for digital cameras capturing text images. For example, indoor illumination differs significantly from outdoor lighting conditions. Finally, the captured images are normally distorted in perspective, since

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

the image plane will not be perfectly parallel to the document plane. The problems are even more difficult if we consider cameras of mobile phones (such as Apple's iPhone camera) that are of lower quality than normal digital photo cameras. It is advantageous to instantly recognize text characters on a mobile phone to avoid transmitting large images over the mobile network. However, the limited power of embedded processors restricts the algorithmic possibilities, since we wish to achieve an acceptable OCR runtime behavior. Currently, OCR systems for mobile phones are available [18, 19], but most of them focus on OCR for specific types of documents like business cards.

In this paper, we present an OCR system for Apple's iPhone that processes different types of text documents in reasonable time. The system is based on Google's freely available OCR engine tesseract [20]. Several image processing and image enhancement techniques are employed to finally improve the OCR result. We have created a test set containing 46 images that consists of several document types: ISO/DINA4 documents and letters, menus, business cards, service manuals, and job advertisements. The impact of the image processing techniques is analyzed in detail on this test set, leading to a recommendation for a practical mobile OCR solution. When using the proposed techniques, the recognition rate of the tesseract baseline OCR system is improved by more than 12%.

The paper is organized as follows. In Section 2, related work is discussed. The proposed OCR system for the iPhone including image enhancement techniques is presented in Section 3. Experimental results for several tests are reported in Section 4. Section 5 concludes the paper and outlines areas for future research.

2 Related Work

Optical character recognition has been a research field for several years, and many solutions have been suggested in the literature. For example, Govindan and Shivaprasad [6] present an overview about character recognition methods in general, while Due Trier et al. [1] provide a survey about feature extraction methods for OCR. Tesseract [18] is freely available software provided by Google that has achieved top-performance in an evaluation study in 1995: in UNLV's (University of Nevada, Las Vegas) OCR evaluation test [15].

With respect to image and video OCR, some work exists that aims at localizing text in pictures and video frames (e.g. [5]), or removing complex image background for improving a subsequent OCR process (e.g. [6]), called text segmentation. Jung et al. [10] review a number of methods for text extraction in images, while Lienhart [11] reviews some methods for video OCR.

With respect to OCR using *mobile* devices for both capturing the image *and* text recognition, there are only few proposals that touch the field of OCR. Iwamura et al. [8] present a real-time system for camera-based recognition of pictograms and characters. It is based on adaptive binarization and contour extraction coupled with a geometric hashing method. They report an accuracy of 50% to 85% for different fonts. However, in their experiments they treat different characters as belonging to the same class, which makes a comparison with other works difficult. Liu et al. [12] present some tools for visually impaired people, including a currency reader. Mollah et al. [13] present a system for extracting text regions from business cards. To the best

of our knowledge, there are no published investigations of a mobile OCR system that also works for classical document types apart from business cards, except for Joshi et al.'s work [9]. They propose a system that uses Tesseract's OCR as well as some preprocessing routines on a backend server, that is the image data have to be transferred to this server, in contrast to our pure phone-based solution.

3 Mobile OCR on the iPhone

In this section, iPhone specific implementation details and relevant image preprocessing operations are presented. The paper mainly focuses on the issue of how to process text images captured with the iPhone camera by an OCR system in reasonable time, while at the same time obtaining good recognition results for different types of text documents. Hence, simple and effective preprocessing operations are proposed and implemented on the iPhone, and it is investigated empirically which preprocessing and image resolution are necessary for successful optical character recognition.

3.1 iPhone Implementation

The proposed iPhone application has been developed using Apple's integrated development environment XCode and the Interface Builder. Apple provides an Objective-C compiler for implementing iPhone applications. This compiler allows developers to mix C++ and Objective-C code. Tesseract version 2.04 has been used in our system. Since tesseract is implemented in C/C++, a static tesseract library can be compiled for the iPhone using an appropriate makefile. When including the tesseract library into the iPhone application, it is necessary to adequately declare the tesseract class *TessBaseAPI* for Objective-C code. Within the tesseract software, the directory *tessdata* contains some important files for tesseract. When it has been inserted in the XCode project, it will be copied on the iPhone to the folder "appname".app and can be accessed by the application. To display the activity icon while the OCR process is running, the recognition process has to be started in a separate thread.

3.2 Image Preprocessing Operations for Mobile OCR

The OCR process consists of image scanning/capturing, image preprocessing, feature extraction, classification, and post-processing [4]. In this section, we address the preprocessing stage and describe several image processing and image enhancement techniques that we expect to improve OCR on images captured with the iPhone. In particular, the techniques address the issues of low resolution images and non-uniform illumination patterns on the document. The techniques comprise resolution enhancement, contrast enhancement, median filtering, blurring, adaptive local binarization, and morphological operations such as opening and closing. The techniques are described in detail below, all techniques have been implemented on the iPhone using the programming language Objective-C.

Resolution Enhancement. Most OCR systems are developed based on the assumption that documents are scanned at least with 300 dots per inch (dpi). The resolution of the iPhone 3GS camera is 1536*2048, for an ISO/DINA4 document this yields a resolution of approximately 180 dpi. To obtain a resolution of 300 dpi for

optical character recognition, the image resolution for an ISO/DINA4 document would have to be 2480*3508, whereas for a business card of size 3.3*2.6 inches only 1004*768 pixels are required for 300 dpi. Hence, image resizing is an important tool for OCR pre-processing, which is realized by linear interpolation in our system. The source pixels are copied to a target image using a scaling factor, normally yielding floating point coordinates for the source pixels in the target. The pixel values at integer positions are computed using the four pixel neighbors that were copied from the source image. They are weighted according to the proximity to the target position in the target image. Let (x_l, y_t), (x_r, y_t), (x_l, y_b), (x_r, y_b) be the positions of the neighboring pixels of pixel (x, y) in the target image (x and y are integers). The intensity value in the target image is computed by:

$$I(x,y) = a \cdot I(x_l, y_t) + b \cdot I(x_l, y_b) + c \cdot I(x_r, y_t) + d \cdot I(x_r, y_b) \quad (1)$$

$$a = 1 - \frac{(x - x_l)\cdot(y - y_t)}{(x_r - x_l)\cdot(y_b - y_t)},\ b = 1 - \frac{(x - x_l)\cdot(y_b - y)}{(x_r - x_l)\cdot(y_b - y_t)}, \quad (2, 3)$$

$$c = 1 - \frac{(x_r - x)\cdot(y - y_t)}{(x_r - x_l)\cdot(y_b - y_t)},\ d = 1 - \frac{(x_r - x)\cdot(y_b - y)}{(x_r - x_l)\cdot(y_b - y_t)}, \quad (4, 5)$$

where a, b, c, and d are the weights for the neighboring pixels.

Contrast Enhancement. The contrast of images captured in an unrestricted outdoor or indoor environment is often not optimal. Often, the range of possible intensity values is not exploited in images, and the contrast is low. Histogram equalization is a known technique [3] for improving the contrast of an image. A new histogram is calculated as follows. The p percent of pixels (e.g., p=1%) called *outliers* with the highest or lowest intensity value are mapped to pixel values of 255 or 0, respectively. This is done to remove the impact of outliers. Let *min* and *max* be the minimum and maximum of the remaining values. Then, the interval of the remaining pixels in the histogram is computed and is mapped to the new histogram that covers the whole range of values [0, 255]:

$$newval = \frac{255}{\max - \min} \cdot (val - \min) \quad (6)$$

The intensity of all pixels with value *val* (with $val \geq min$ and $val \leq max$) in the original images is replaced by *newval* in the destination image, pixel intensities of outlier values $val < min$ or $val > max$ are set to 0 and 255, respectively.

Median Filtering. To filter out noisy pixels, we apply a median filter of size 3*3 to the image. A filter of size 3*3 (or 5*5) is shifted over the image and the respective center pixel and its 8 (24) neighbors are taken into account. Then, these pixel values are sorted and the median value, i.e. the value at middle position 5 (13) in the sorted list, is copied to the corresponding pixel position in the target image.

Adaptive Local Binarization. The illumination of the image cannot be expected to be uniform for arbitrarily captured images. Hence, global thresholding will not achieve optimal segmentation results. Two alternative local binarization methods for improving OCR results are investigated. The first method applies a local threshold that equals the average value in a region of size n*n, while the second method uses Otsu's method [14] to estimate a threshold: This method aims at finding a threshold that separates the histogram in two classes and minimizes the intra-class variance and maximizes the inter-class variance.

Smoothness/Blur Filter. To smooth the image, a filter matrix of size *n*n* is shifted over the image. We have defined two different filters M_1 and M_2 for our experiments.

$$M_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 5 & 5 & 5 & 1 \\ 1 & 5 & 56 & 5 & 1 \\ 1 & 5 & 5 & 5 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad M_2 = \begin{pmatrix} 1 & 3 & 4 & 3 & 1 \\ 3 & 12 & 20 & 12 & 3 \\ 4 & 20 & 56 & 20 & 4 \\ 3 & 12 & 20 & 12 & 3 \\ 1 & 3 & 4 & 3 & 1 \end{pmatrix} \tag{7, 8}$$

Morphological Operations. Morphological filters can be used to react to certain structures in binary images. To enhance an image for subsequent OCR, it can be useful to remove noisy image pixels, or to grow or shrink the thickness of characters. Erosion and dilation are morphological operations that apply a structure element to an image. For example, the structure element can define the 4- or 8-neighborhood of a pixel. In case of dilation, the structure element is connected with an image region according to an OR-relation, i.e. a black pixel is set at the current pixel position if the center *or one* of the neighbor pixels is black; in case of erosion, it is connected according to an AND-relation, i.e. a black pixel is set if the center pixel *and all* neighbor pixels are black. The operations erosion and dilation can be used to realize open (erosion followed by dilation) or close (dilation followed by erosion) operations by consecutive execution of erosion and dilation.

5 Experimental Results

In this section, we present experimental results for a test set of 46 images that were captured using the iPhone. The test set covers a variety of document types covering different document types (Figure 1): ISO/DINA4 documents, letters, menus, business cards, service manuals, and job advertisements. In a comprehensive experimental test setup consisting of more than 200 experiments, the impact of different pre-processing techniques is investigated systematically. Finally, the performance is compared with two commercial iPhone OCR systems.

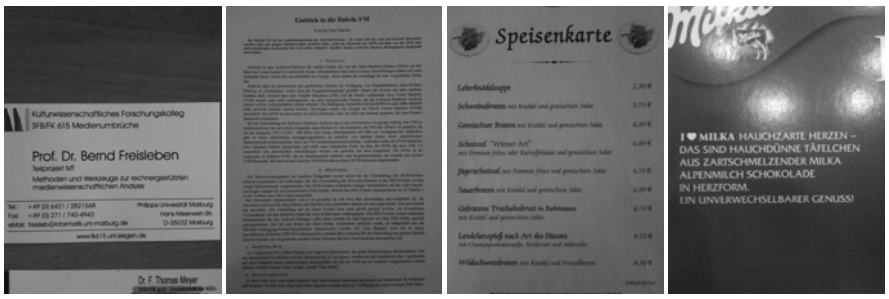


Figure 1: Examples of the test set; all images were captured with the iPhone.

The results are presented in Table 1-5 in terms of accuracy of character recognition. First, the impact of different image resolutions is tested in conjunction with two global thresholding methods (Table 1). Table 2 shows results for experiments investigating the impact of adaptive local thresholding. Further experiments are conducted using different region sizes for local thresholding,

histogram equalization, median filtering, blur filtering, and morphological operations. It turns out that the resolution of the iPhone images is sufficient to obtain satisfying results (Table 3, rank 9), although Table 3 shows that best results are obtained when resizing the image to larger resolutions, using blur or median filtering, histogram equalization for improved contrast, and using morphological operations. The most important thing to significantly improve the results of the tesseract OCR engine is to use adaptive local thresholding. The simple binarization method based on the region average was more efficient than Otsu's method that did not give satisfying results in dark regions at the image border. A comparison of the tesseract baseline system, our best parameter setting for improving OCR and two commercial mobile iPhone OCR systems (Babelshot [18] and PerfectOCR [19]) is presented in Table 4. The system using a resolution of 2000*2666 pixels, median filtering, local thresholds for regions of size 100*100 pixels, and finally using a blur filter yielded an accuracy of 74.9% and outperformed both commercial systems by 11.5% and 5.7%, respectively. The runtime differed with respect to the number of characters in an image and the used image processing operations. Using our best system, processing a business card took about 15-25 sec., while processing an ISO/DINA4 text document took 65-220 seconds. Runtimes are presented in Table 5 for some image examples using the proposed system and both commercial iPhone OCR systems.

Figure 2: Binarization result using local thresholding, region size is 350*350.

5 Conclusions

In this paper, we have presented a mobile iPhone OCR system that is able to process different types of text documents that were captured by the iPhone camera. The system is based on Google's freely available OCR engine tesseract. The input image is enhanced by adequate pre-processing to improve the final OCR result. We have tested the impact of different image processing and image enhancement techniques on a test set consisting of different document types: ISO/DINA4 documents and letters, menus, business cards, service manuals, and job advertisements. The recognition rate of the tesseract baseline OCR system using the proposed techniques could be improved by more than 12%, outperforming two commercial iPhone applications, while at the same time achieving an acceptable runtime on the mobile phone.

In the future it is planned to port the software to other mobile platforms, such as Google's Android or Windows Mobile. Finally, the runtime should be reduced by optimizing the code for the target platforms, possibly by using assembly language.

Table 1: Results for baseline OCR system using different image resolutions.

Accuracy [%]	Baseline	Threshold global	Otsu global
Image size 1536*2048 (orig.)	62.7	**65.6**	62.5
Image size 2000*2666	60.9	64.9	60.6
Image size 2550*3315	59.8	62.9	59.7

Table 2: Results for three region sizes (in brackets) for local thresholding.

Accuracy [%]	Local threshold (50)	Local threshold (150)	Local threshold (500)	Otsu (50)	Otsu (200)	Otsu (500)
1536*2048	73.8	72.0	69.8	47.8	57.6	66.9
2000*2666	73.3	73.9	70.0	45.6	51.6	63.8
2550*3315	69.0	71.2	70.7	47.8	50.0	63.4

Table 3: The parameter settings of the top 10 results of 207 tests.

Rank	Preprocessing steps in terms of order	Accuracy
1	2000*2666, Med. (3), Local Thresh (100), Blur (M_1)	**74.9%**
2	2000*2666, Med. (3), Local Thresh (100), Open (4), Blur(M_1)	74.2%
3	2550*3315, Contrast, Blur (M_2), Local Thresh (100), Close (4), Med. (3)	74.1%
4	2000*2666, Med. (3), Local Thresh (100), Close (8), Blur(M_1)	74.0%
5	2000*2666, Med. (3), Local Thresh (100), Close (8), Open (8), Blur (M_1)	74.0%
6	2000*2666, Local Thresh (150)	73.9%
7	2000*2666, Local Thresh (100)	73.9%
8	2550*3315, Contrast, Blur (M_2), Local Thresh (100), Open (4), Med. (3)	73.8%
9	1536*2048 (No Resizing), Local Thresh (50)	73.8%
10	2550*3315, Contrast, Blur (M_2), Local Thresh (100), Med. (3), Open (4)	73.7%

Table 4: Comparison with two commercial mobile iPhone OCR systems.

Accuracy [%]	Babelshot	PerfectOCR	Tesseract	Proposed (best)
Character recognition	63.4	69.2	62.7	**74.9**

Table 5: Runtimes of two commercial iPhone OCR systems and our approach.

Runtime [sec]	DINA4 Image 1	DINA4 Image 2	Short Text 1	Short Text 2	Business Card 1	Business Card 2	**Avg.**
Babelshot	129	87	9	14	8	15	**44**
Perfect OCR	75	186	31	25	22	20	**60**
Proposed System	67	127	17	21	15	23	**45**

References

1. Arica, N. and Yarman-Vural, F.T.: An Overview of Character Recognition Focused on Off-linehandwriting. In IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, May 2001, vol. 31, Issue 2, 216-233.

2. Ball, G. and Srihari, S.: Semi-supervised Learning for Handwriting Recognition. In Proceedings of 10th International Conference on Document Analysis and Receognition, Barcelona, Spain, 2009, 26-30.
3. Burger, W. and Burge, M.J.: Digital Image Processing: An Algorithmic Introduction using Java. Springer, 2008.
4. Due Trier, O., Jain, A.K., Taxt, T.: Feature Extraction Methods for Character Recognition-A Survey. In *Pattern Recognition (29)*, Elsevier, No. 4, 1996, 641-662.
5. Gllavata, J., Ewerth, R., Freisleben, B.: Text Detection in Images Based on Unsupervised Classification of High-Frequency Wavelet Coefficients. In Proceedings of 17th Int'l Conf. on Pattern Recognition, Vol. 1, Cambridge, UK, 2004, 425-428.
6. Gllavata, J., Ewerth, R., Stefi, T., and Freisleben, B. Unsupervised Text Segmentation Using Color and Wavelet Features. In Proceedings of the 3rd International Conference on Image and Video Retrieval 2004, Lecture Notes on Computer Science LNCS 3115, Dublin, Springer, 2004, 216-224.
7. Govindan, V.K. and Shivaprasad, A.P.: Character Recognition - A Review. In *Pattern Recognition*, Elsevier, Volume 23, No. 7, 1990, 671-683.
8. Iwamura, M., Tsuji, T., Horimatsu, A., and Kise, K.: Real-Time Camera-Based Recognition of Characters and Pictograms. In Proceedings of 10th International Conference on Document Analysis and Recognition, Barcelona, Spain, 2009, 76-80.
9. Joshi, A., Zhang, M., Kadmawala, R., Dantu, K., Poduri, S., and Sukhatme, G.: OCRdroid: A Framework to Digitize Text Using Mobile Phone. In Proceedings of ICST International Conference on Mobile Computing, Applications, and Services. San Diego, CA, USA, 2009, online available at: http://cres.usc.edu/cgi-bin/print_pub_details.pl?pubid=635.
10. Jung, K., Kim, K.I., Jain. A.K.: Text Information Extraction in Images and Video: A Survey. In Pattern Recognition, Volume 37, Issue 5, 2004, 977-997.
11. Lienhart, R.: Video OCR: A Survey and Practitioner's Guide. In Video Mining, Kluwer Academic Publisher, Oct. 2003, 155-184.
12. Liu, X., Doermann, D., and Li, H.: Mobile Visual Aid Tools for Users with Visual Impairments. In Lecture Notes in Computer Science (LNCS 5960), Mobile Multimedia Processing, Springer, 2009, 21-36.
13. Mollah, A.F., Bas, S., Das, N., Sarkar, R., Nasipuri, M., Kundu, M.: Text Region Extraction from Business Card Images for Mobile Devices. In Proceedings of International Conference on Information Technology and Business Intelligence, Nagpur, India, 2009, 227-235.
14. Otsu, N.: A Threshold Selection Method from Grey Level Histograms. In IEEE Transactions on Systems, Man, and Cybernetics, 9, 1979, 62-66.
15. Rice, S.V., Jenkins, F.R., and Nartker, T.A.: The Fourth Annual Test of OCR Accuracy. Available online at: http://www.isri.unlv.edu/downloads/AT-1995.pdf
16. Smith, R.: An Overview of the Tesseract OCR Engine. In Proceedings of 10th International Conference on Document Analysis and Recognition 2007, Curitiba, Paraná, Brazil, 2007, 629-633.
17. Sundaram, S. and Ramakrishnan, A.G.: An Improved Online Tamil Character Recognition Engine using Post-Processing Methods. In Proc. of 10th Int'l Conference on Document Analysis and Recognition, Barcelona, Spain, 2009, 1216-1220.
18. Babelshot (Photo Translator). Available online at: http://itunes.apple.com/us/app/babelshot-photo-translator/id334194705?mt=8
19. PerfectOCR: Document scanner with high quality OCR. Available online at: http://itunes.apple.com/de/app/perfect-ocr-document-scanner/id363095388?mt=8
20. Tesseract: tesseract-ocr: Available online at: http://code.google.com/p/tesseract-ocr/

Application of RFID Technology in Management of Controlled Drugs- Proof of Concept

Yuan-Nian Hsu[1,4], Shou-Wei Chien[4], Vincent Tsu-Hsin Lin[5], Jimmy Cheng-Ming Li[5,6], Tan-Hsu Tan[7], Yung-Fu Chen[2,3,8,*]

[1]Department of Health Care Administration, [2]Department of Management Information Systems, and [3]Institute of Biomedical Engineering and Material Science, Central Taiwan University of Science and Technology, Taichung, Taiwan
[4]Taichung Hospital, Department of Health, Executive Yuan, Taichung, Taiwan
[5]Institute for Information Industry, Taipei, Taiwan
[6]Department of Mechanical Engineering and [7]Department of Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan
[8]Department of Health Services Administration, China Medical University, Taichung, Taiwan
{taic, sowejan}@mail.taic.doh.gov.tw, {vincelin,jimmyli}@iii.org.tw, thtan@ntut.edu.tw, yungfu@mail.cmu.edu.tw

Abstract. Due to their potential for habitual use, dependence, abuse, and danger to the society, controlled drugs are tightly regulated in many countries. With the integration of radio frequency identification (RFID) system, controlled drug management can be more accurate and effective. Furthermore, RFID could be the key role for the discrimination between genuine and counterfeit drugs. Hence, how to elevate patient safety is becoming a very important issue for healthcare organizations. Medication safety is the most important topic in improving patient safety. In addition to reducing the error rate of records written manually in the traditional procedures, the application of RFID technology is even more effective to control and monitor legal disposition of controlled drugs and to ensure the accuracy of their pedigrees.

Keywords: Radio Frequency Identification (RFID), Medication Error, Patient Safety, Controlled Drugs

1 Introduction

Current regulations require that the manufacture, prescription, and administration of controlled drugs have to be declared periodically. In Taiwan, pharmaceutical manufacturers have to report the manufactured quantity monthly and the healthcare organizations have to report the prescribed and stocked quantity every 6 months. These declarations have to be as precise as to be counted in the smallest unit. For examples, the declaration unit for Pastils should be in "tablet" and the unit for drugs in liquid type should be "ml" or "cc". In the current regulation of Taiwan, the auditors

*Corresponding author: No. 91, Hseuh-Shih Road, Taichung 40402, Taiwan, R.O.C. Tel:+886-4-22053366 Ext. 6315, Fax:+886-4- 22031108

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

of National Bureau of Controlled Drugs (NBCD) visit individual organizations regularly for an audition of written records to ensure that the recorded quantity of production is consistent to the prescribed dosages plus available stocks. However, the disposition of the controlled drugs around the country could not be known in real time and, therefore, difficult to control and manage.

The aim of this study was to apply the RFID technology to offer a unique ID to each controlled drug during the manufacturing stage. This unique ID was linked with the tag ID stuck on the package box and stored in the pedigree database for later reference. During the manufacturing and dispensing procedures, any status change will be automatically detected and recorded by RFID readers. In addition to facilitating logistic operation and reducing human intervention, RFID may also enable the medical staffs to automatically acquire the status of drugs at various critical checkpoints based on their IDs during administrating, dispensing and stocking, thereby speeding up the identification and verification processes. Furthermore, the consumed quantity will be automatically recorded and updated in the pedigree database to simplify the declaration procedure. Concerning the pharmaceutical management, the disposition and flow control may be updated in a real-time manner. For a batch of drugs that have caused adverse drug events, the recalling procedures and remedial operations in the aftermath can be facilitated through manipulation of the pedigree database.

1.1 Medication Adverse Events

It was reported that the annual mortality for adverse events was between 44,000 to 98,000 in US. Meanwhile, the cost for preventing the occurrence of adverse events was as high as 17-29 billion US dollars, accounting for over 50% of the total medical cost [1]. Furthermore, medication error has been the most common malpractices occurred in healthcare organizations.

As estimated by World Health Organization (WHO), the annual trade of counterfeit drugs approximates 35 billion US dollars [2]. Most people cannot easily discriminate genuine from the counterfeit drugs. The application of RFID has become one of the most effective ways to cope with this problem. According to WHO, the application of RFID may reduce the counterfeit drug rate by 30% and 6-10% for developing countries and developed countries, respectively.

According to US Pharmacopeia-Medication Errors Reporting Program (USP-MERP), over one third of medication errors were related to similarity of drug names and such confusion could happen for both generic names and trade names. Labeling and packaging is the second factor causing medication errors. Since most medications produced by a pharmaceutical factory are generally with the same design of labels and packages [3], the error rate will increase if numerous medications are simultaneously purchased from one single pharmaceutical factory. The risk of confusion will be even higher if the solution of two different medications having similar color is packaged into ampoules or vials with the same size, shape, and color. Similar serious consequences will occur if such errors happened for controlled drugs.

US Food and Drug Administration (FDA) has started adopting RFID technology since 2004 and promoted anti-counterfeit drug program progressively from 2004 to

2007 in 3 stages. During the first year, evaluation of feasibility was conducted. In the following year, RFID tags had been attached on the pallets and outer package of high-risk products, which were then extended to other products later this year. In 2007, RFID tags were overall introduced to the pallets and outer boxes for all products, as well as the outer package for most products [4].

1.2 Management of Controlled Drugs

Controlled drugs are defined as addictive narcotics, psychotropic drugs, and other drugs requiring a stricter control [5]. Based on whether they are habit-forming and the extent to which they are leading to dependence, abuse, and social hazards, controlled drugs are classified into 4 schedules for efficient management and control. Controlled drugs, for example opioid analgesics, are generally used to control pain caused by cancer, trauma, surgery, and other conditions [6-8]. Non-medical use and abuse accompanied with an increase of prescription rate of controlled drugs have raised a public health problem, especially for young adults with age ranging from 18 to 24 years old [9].

The export, import, and manufacture of Schedules 1 and 2 drugs are managed by the pharmaceutical plant of NBCD in Taiwan with an annual production of 3.6 million injections and import of 0.3 million doses of various medications. The NBCD plant also manufactures drugs in bottles and tablets of PTP (press through package). Imported controlled drugs are mainly Fentanyl patches. The drugs of Schedules 3 and 4 are mainly Ketamine, Flunitrazepam, Secobarbital and Amobarbital.

The management of controlled drugs is based on the Controlled Drug Act, which was stipulated according to the Single Convention on Narcotic Drugs of 1961, the Convention on Psychotropic Substances of 1971, and United Nations Convention against Illicit Traffic in Narcotic Drugs and Psychotropic Substances of 1988. The essences of managing controlled drugs include classification, licensing policy, disposition, and legitimated use managements.

Inappropriate use of controlled drugs may lead to addiction and improper management may cause illicit use. However, these drugs are necessary and inevitable for a proper administration for alleviating and treating certain diseases. In Taiwan, controlled drugs-related regulatory provisions have been stipulated by Department of Health (DOH) to prevent iatrogenic addiction and provide as guidelines for medical management and prescription. The purpose of these controlling strategies is to use them when necessary and to spare them when inappropriate.

1.3 The application of RFID technology in healthcare industry

In Taiwan, the first application of RFID technology in healthcare industry was to monitor patients and medical staffs entering and leaving quarantine areas during the SARS (Severe Acute Respiratory Syndrome) outbreak period to prevent disease spreading [10]. According to previous studies, the implementation of RFID in operation rooms or intensive care units [11], together with RFID labels on blood bag, medication container and patients' wrists, were shown to be able to facilitate the

management of transfusion and injection agents to effectively ensure patient safety [12]. RFID technology may also be applied to monitor trachea intubating to prevent serious injuries from inappropriate intubation, such as incorrect location and improper depth of intubations [13]. With a correct implementation of RFID tag, RFID reader may precisely monitor if the intubation is at the right and appropriate site, especially during the patient has been transferred to other hospitals or changing his/her posture or position.

2 System Development and Verification

The organizations which were invited to participate in this study included RFID manufacturers and users. In addition to the pharmaceutical plant of NBCD and an affiliated hospital of DOH, a private pharmaceutical enterprises was also asked to provide the testing sites of drug production lines. Together with a simulation procedure of automatic labeling, this study also verified all the necessary procedures needed to be fulfilled by the pharmaceutical industry.

2.1 System Implementation

The scenarios of application of RFID in managing controlled drugs are shown in Fig. 1 and described below:

Pharmaceutical plants: During the manufacturing procedure, a RFID tag was stuck on the body of each bottle and the packaging box to enable on-line and off-line readings automatically. With a reader installed on the gate for online reading, the drug information was uploaded to the DOH database during stocking and distributing operations. The drugs were then distributed to the Sale Department of NBCD. Data collected at various stages were stored in the proof of concept (POC) verification system.

Hospitals: The procedures conducted in hospitals include stocking and withdrawing. Using portable RFID readers, medical staffs are able to manage the stocks of various controlled drugs and upload their related information to the DOH system.

Pedigree database system: Pedigree database system is connected to the POC verification system for updating related information in real time manner.

Declaration system at DOH: This system receives information including drug code, lot number, reason and date of drug stocking/withdrawing, and quantity, of controlled drugs from all processing steps.

POC verification system: The information recorded in this system includes drug codes, lot numbers, manufacture date, and tag ID recorded in the manufacturing stage; tag ID stuck on the packaging box and its matching information of the RFID tags stuck on bottles inside the packaging box which were recorded in the packaging stage; and categories and codes of medications and their pharmaceutical manufacturers recorded at the stocking and distributing stages.

The operation and management of controlled drugs are mainly taken place in 4

sites, including NBCD, sale department of NBCD, medication warehouse of the hospital, and RFID-based storage cabinet at the pharmacy department. The operational procedures are illustrated in Fig. 2 and described below.

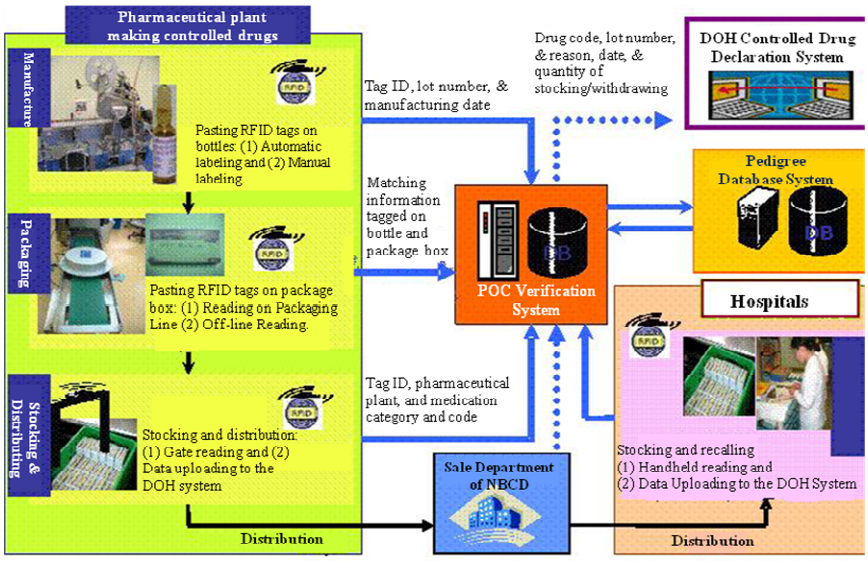


Figure. 1. Scenarios of RFID application in managing controlled drugs .

Pharmaceutical plant: During the manufacturing procedure, RFID tags are automatically stuck on the produced drugs and then packaged into an outer packaging box stuck with an additional RFID tag. For quality control of controlled drugs, produced and packaged injections and PTP tablets are all manually weighted in NBCD pharmaceutical plant to ensure correct dosage. Afterwards, products are transferred to warehouse for storage. Upon receiving the orders from healthcare organizations, requested amount of controlled drugs are transferred to the storeroom of the Sale Department.

NBCD Sale Department: All the drugs stocked at the NBCD Sale Department are all authorized with permission for prescribing. Upon replenishing from the Pharmaceutical plant, RFID tags are read to update the inventory automatically. When a hospital representative personally comes to Sale Department to claim drugs, one should submit the hospital's licensing certificate and the superintendent's seal for inspection and confirmation before receiving the controlled drugs. After checking the drug names, lot numbers, and quantity on the list, the hospital representative has to sign on the list and leave his/her ID number for future references.

Drug warehouse in hospital: During replenishment, RFID tags of controlled drugs will be read and verified before updating pharmaceutical database. To administrate the controlled drugs, a physician has to make a special prescription before dispensing by pharmacists. Both the pharmacist who dispenses the drugs and the nurse who receives

the drugs are asked to sign on the order. Meanwhile, pharmacists have to record each dispensing on the "stocking/withdrawing log of controlled drugs" on a daily basis.

RFID-based storage cabinet: Controlled drugs should be checked carefully according to the items listed on the medication list before moving to the RFID-based storage cabinet for stocking. At the mean time, information of the stocked drugs will be used to update the pharmacy database. Only the authorized pharmacists are allow to open the storage cabinet.

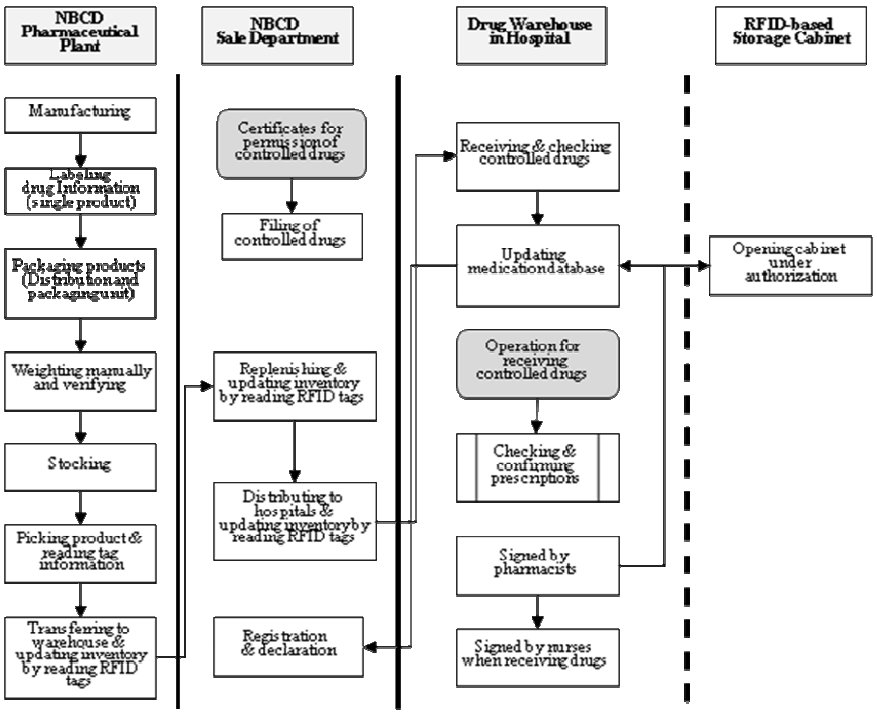


Figure 2: Operation procedures of controlled drugs

2.2 Verification Requirements

In addition to those of general medications, the management and handling of controlled drugs particularly emphasize on the following requirements: (1) Anti-Counterfeiting: validity and legitimacy of the drug sources must be ensured; (2) Targeted Recalls: the scenario for unit-dose recalling must be considered; (3) Product Packaging Authentication: legality and validity of repackaging must be ensured; (4) Receipt Confirmation: follow-up procedures for inconsistent ordering and receiving items must be included; (5) Returns Tracking: returned drugs according to the client's returning procedures must be tracked; (6) Product Visibility: visibility of drug information during individual operational procedures must be assured.

In addition, the tracking functions required for each operational procedure, limitations during the development stage should also be taken into consideration. The verification system must seamlessly integrate various organizations with different operational and managing procedures.

2.3 Verification Index

System verification include basic function test and Scenario verification. Basic functional test is only aimed at testing efficiency of RFID hardware components. The tested items include reader, antenna, and tags. Hardware equipment used for testing the basic function include fixed and handheld RFID readers. Three fixed readers, including UHF Readers (Brand A Gen-2) equipped with far-field and near-field antenna, UHF Reader (Brand B), and HF Reader (Brand A), and two handheld readers, including UHF Reader (Brand A) and HF Reader (Brand B), were tested. Table 1 shows the items used for testing the basic functions.

Table 1. Test items used for testing basic functions and their verification indices

Item	Appearance	Name	Package	Verification Index
Ampoule (1ml)		Morphine Hydrochloride Injection	10 ampoules /box	Reading distance, interfere from liquid inside the bottle, interfere from curvy surface adherence, reading of closely arranged tags, linkage of tags on outer box and tags on a single ampoule
Ampoule (2ml)		Fentanyl Injection	10 ampoules /box	
Aluminum patch		Fentanyl transdermal patch	5 patches/box	Reading distance, effect of metal materials, effect of tag attachment site
PTP aluminum foil		Morphine Sulfate Film Coated Tablets	10 tablets/row, 5 rows/box	Reading distance, effect of metal materials, identification procedure for receiving single tablet
Glass bottle		Codeine Phosphate Tablets	100 tablets /bottle	Reading distance, identification procedure for receiving single tablet

3 System Verification and Evaluation

3.1 Basic function test

The testing results of are the items shown in Table 1 are summarized below:
Determination of reading distance: The reading distance of a large HF tag is longer than that of a small HF tag. Far-Field tag is applicable for the reading with Fixed UHF Reader Brand A plus Far-Field Antenna or Fixed UHF Reader Brand B. Near-field tag is suitable for reading with Fixed UHF Reader Brand A plus Near-Field Antenna. Far/Near-Field tag could be used for reading with Fixed UHF Reader Brand A plus Far-Field Antenna or Near-Field Antenna. For the choice of antenna for collocation, the condition of on-site reading should be taken into consideration.

Massive reading test: The feasibility analysis of massive reading (10 tags were reading simultaneously) was conducted by massive reading of 1 ml- and 2 ml-ampoules, respectively, packed in a wave box. The ampoules were filled with water to replace real drugs for simulation. During the reading test, times of the wave box passing the antenna of a reader were not limited. It was found that all the readers have the capability of reading 10 tags simultaneously. However, the anti-collision function of HF reader is slightly weaker that a little more reading times (5 times in average) were needed to completely gather information of 10 tags. In contrast, the required reading times of UHF readers and tags are less than that of HF reader (2.5 times in average). Furthermore, the overall reading of Tag H (UHF Tags 29 mm × 13 mm) achieves the best efficiency. The feasibility of HF reader and tags may be promoted if the data processing system may be modified to transmit the HF tag information to database instantly after reading.

Successful reading rates against attaching sites: According to the outcomes of reading distance, reading range, and massive reading test, the tags adopted for further tests included Tag G (UHF Tags 42 mm × 22 mm), Tag H (UHF Tags 29 mm × 13 mm), Tag I (UHF Tags 24 mm × 24 mm), Tag J (UHF Tags 16 mm × 8 mm), and Tag K (UHF Tags 9 mm × 9 mm). There are several reasons that the RFID tag is not suitable to be attached at the bottle bottom. First, the bottle bottom is not flat but with a concave curve, it is not easy to attach a tag on the bottom of a bottle. In this case, a tag is very easily to detach when loosely pasted due to an insufficient attachment surface. Second, it is against the current procedure of tag pasting in the production line, and is unfavorable to change the operational procedure. Third, pasting tags at the bottle bottom is unfavorable for reading tags attached on ampoules packed inside a wave box, since the UHF antenna is generally installed above the conveying belt for a production line. Tags attached at the bottle bottom are 90 degree to the antenna, which is not suitable for reading. Hence, attachment of RFID tags on the bottle body is more beneficial to integrate with the current available text tags which provide visible information of the drugs.

Successful reading rate against moving speed: For a transporting speed below 20 m/min along with a distance of 10-15 mm between antenna and wave box, the successful reading rate is over 90% for all types of tags.

Regarding the stability and successful rate of overall readings, H Tags (UHF Tags 29 mm × 13 mm) demonstrate to be the best choice. The tag stuck on the packaging box and tags attached on the ampoules inside can be read simultaneously. Information embedded in two kinds of tags can be correlated at the backend system. The overall evaluated outcomes of the basic function test is concluded as follows:

(1) Small items, such as ampoules, stuck with Near-Field Tags or Far/Near Field Tags read with Fixed UHF Reader Brand A plus Near-Field Antenna tend to have higher correct rate. Long-distance reading is not required, but should ensure high successful reading rate.

(2) Large items, such as packaging boxes and distribution carton, stuck with Far-Field Tag are suitable to be read with Fixed UHF Reader Brand A or Fixed UHF Reader Brand B plus Far-Field Antenna. Longer reading distance as well as larger reading capacity are required for efficient reading. This tag should be correctly associated with the tags inside the package.

3.2 Scenario Verification

The feasibility study for RFID tags with auto-labeling using labeling machine: In pharmaceutical plants, the present procedure for drug labeling is accomplished with automatic labeling machines. Since various models of labeling machines have been designed and employed for labeling different types of products, this study only tested the feasibility of labeling machine and RFID tags for ampoules to determine their tag damage rate and reading rate. The results are concluded as follows:

(1) During the manual packaging of tags into tag rollers, inappropriate operation may cause a slight damage of tags leading to a reduced overall reading efficiency. Such problem could be avoided if RFID tags used for automatic labeling machine could be produced in roller form.

(2) During the labeling procedure, ampoules are turned back to the ampoule turntable via guardrails at the side of collection plate. This procedure caused the breakage and damage of the hinge linking tag chips and antenna. To reduce the tag damage rate, the material of guardrails at the side of collection plate should be modified (such as rubber). The other way is to design and build a fillister at guardrails to avoid a direct collision between ampoules and chips to protect tags from collision-caused damage.

(3) During labeling procedure, static electricity could be generated by the friction of ampoule rolling actuator of labeling machines and the transmission of generated static electricity to tag chips may lead to chip damages. The installment of ground wire in this area is recommended to guide away possible static electricity to protect chips. The material of ampoule rolling actuator could be replaced with material which is static electricity-proof to avoid the generation of static electricity.

(4) The overall reliability of automatic labeling with current labeling machine and commercial RFID tags is lower than 90%. Without an appropriate modification of automatic labeling machine, damage rate of RFID tag could not be effectively improved.

(5) Since the mechanical modification of automatic labeling machine was not in the scope of this POC verification, the integration of labeling machine and RFID tag labeling was not tested at field verification.

RFID tag damage rate and reading rate: The RFID tags still function properly after automatic labeling were subjected to reading tests. After being placed in a wave box, ampoule was read with Fixed UHF Reader Brand A plus Near-Field Antenna resulting in the following conclusions:

(1) Upon the effective prevention of permanent damage to RFID tags caused by automatic labeling machine during auto-labeling procedure, RFID tags may effectively collocate with automatic labeling machines.

(2) With auto-labeling, ampoules packaged in a wave box may be directly read via Fixed UHF Reader Brand A plus Near-Field Antenna.

Successful reading rate for wave boxes on the transporting belts: This reading test was conducted using ampoules transporting belt in the pharmaceutical plant. The tested items were ampoules manually labeled at the Stage 1 and ampoules labeled

automatically by labeling machines. Reading was conducted while these ampoules were packaged inside the wave box. Ampoules transporting belt is made of steel skeleton with plastic belt and therefore, Near-Field Antenna is installed above the transporting belt. Reading at this point includes all ampoule tags inside the box and the corresponding tag on the outer wave box. The testing results are as summarized below:

(1) By testing various ways for wave box passing the reader, the overall successful reading rate for ampoules passing the reader consecutively is the best. The reason could be that it may create a time difference between any two tags and therefore reader has no need to process mass amount of tag information at the same time, which eventually enhances the successful reading rate.

(2) There is only slight difference in overall reading efficiencies between manual labeling and automatic labeling. However, tags labeled automaticly have to be checked if their functions are normal before reading tests.

Reading test of RFID tags thermally damaged by shrink film machines: This test is to determine if working temperature will cause any damage of RFID tags. The heat source of tested shrink film machine is at the center with a temperature of 170℃ and the heating way is radial heating. The duration of medications passing through this machine is about 5 seconds. Test results are as follow:

(1) After being packaged and heated through shrink film machine, all tags could still be accurately read.

(2) Commercially available readers and RFID tags all could be applied in the packaging procedure with shrink film machine.

3.3 Field Verification

This verification has shown that the introduction of RFID into the supply chain of controlled drugs may improve the operation efficiency. This improvement is mostly the outcome of changing original manual operation to automatic operation assisted by RFID technology. Together with automatic data extraction from the system, medication safety is highly ensured with the benefits and operational efficiency summarized in Tables 2 and 3.

4 Conclusion

The application of RFID technology in the supply chain of controlled drugs provides management efficiency. In this study, various tags of Near-Field, Far-Field, and Near/Far-Field, as well as readers have been evaluated to determine applicable packaging and settings for controlled drugs. Furthermore, it has been confirmed that application of RFID technology integrated with automatic data extraction in the manufacturing and transporting procedures may lead to an easier and more accurate medication pedigree. This POC verification only aimed at the technical feasibility and preliminary efficiency. The future study will aim at the application of RFID in various

operation steps. The joint participation of private enterprises will be encouraged. Furthermore, the feasibility of applying RFID technology in establishing a comprehensive mediation pedigree will also be investigated in the future.

Table 2. Benefits of RFID applied in hospital

Procedure	Current procedure	RFID-based procedure
Receive/check before stocking	Visual comparison of documents	Auto data extraction and mass reading to speed up the verification and avoid human errors
Inventory	Visual comparison of documents	Auto data extraction and mass reading to speed up the verification and avoid human errors
Dispensing verification	Visual comparison of prescriptions	Auto data extraction and prescription verification to avoid human error
Withdraw/receive recording	It needs 2-3 hours/day to input data with MS Excel	Automatic data extraction and registration eliminate the need for manual entry to avoid human errors
Declaration	An integrated declaration for every 6 months	A real-time declaration

Table 3. Benefits of RFID applied in pharmaceutical plants and sale department

Procedure	Current procedure	RFID-based procedure
Drug Distribution	Not easy to control and could only be traced back to lot No.	With a unique ID, each product could be traced as long as the record is stored in the system.
Tablets in foil	With a daily production of 130000 tablets (2600 boxes/day), 8 hours/day is required for manual weighting.	With automatic data extraction, auditing, and accumulation of product quantity, it can enhance the processing speed and avoid manual errors.
Injection	With a daily production of 30,000 injections (3,000 boxes/day), 6 hour/day is needed for manual weighting	With automatic data extraction, auditing, and accumulation of product quantity, it can enhance the processing speed and avoid manual errors.
Check/ receive when stocking	Visual comparison of documentation	With mass data reading, it can speed up the inspection procedures and avoid manual errors.
Inventory	Visual comparison	With mass data reading, it can speed up the inspection procedures and avoid manual errors.
Distribution	Visual comparison of documentation	With an automatic data extraction and comparison, it can speed up the inspection procedures and avoid manual errors.
Declaration	Monthly declaration	Real-time declaration

Acknowledgments: The study was partially supported by Institute for Information Industry (2007 RFID POC project) and National Science Council (Grant No. NSC98-2410-H-039-003-MY2) of Taiwan.

References

1. Kohn, L.T., Corrigan, J.M., and Donaldson, M.S.: To Err is Human: Building a Safer Health System, National Academy Press, Washington DC, April 2000
2. World Health Organization , http://www.who.int/en/
3. Hoffman, J.M., and Proulx, S.M.: Medication Errors Caused by Confusion of Drug Names. Drug Safety 26(7), 445-452 (2003)
4. Food and Drug Administration (FDA): Radiofrequency Identification Feasibility Studies and Pilot Programs for Drugs. http://www.fda.gov/oc/initiatives/counterfeit/rfid_cpg.html (2004)

5. Department of Health, Executive Yuan, Taiwan, R.O.C., http://www.nbcd.gov.tw/home/control/content01.aspx
6. Joranson, D. E., Ryan, K.M., Gilson, A.M., et al.: Trends in Medical Use and Abuse of Opioid Analgesics. JAMA 283(13), 1710-1714 (2000)
7. Brawley, O.W., Smith, D.E., and Kirch, R.A.: Taking Action to Ease Suffering: Advancing Cancer Pain Control as a Health Care Priority. CA Cancer J Clin 59, 285-289 (2009)
8. Dunn, K.M., Saunders, K.W., Rutter, C.M., Banta-Green, C.J., Merrill, J.O., Sullivan, M.D., Weisner, C.M., Silverberg, M.J., Campbell, C.I., Psaty, B.M., and Von Korff, M.: Opioid Prescriptions for Chronic Pain and Overdose: A Cohort Study. Ann Intern Med 152, 85-92 (2010)
9. McCabe, S.E.: Screening for Drug Abuse Among Medical and Nonmedical Users of Prescription Drugs in a Probability Sample of College Students. Arch Pediatr Adolesc Med 162(3), 225-231 (2008)
10. Science & Technology Policy Research and Information Center, N., Application of RFID in Healthcare Setting, from http://cdnet.stpi.org.tw/techroom/market/eerfid/
11. Chen, P.J., Chen, Y.F., Chai, S.K., Huang, Y.F.: Implementation of an RFID-based management system for operation room. Proc. of 8th ICMLC, Baoding, China, July, 2009
12. Dzik, S.: Radio Frequency Identification for Prevention of Bedside Errors. Transfusion 47, 125S-129S (2007)
13. Reicher, J., Reicher, D., and Reicher, M.: Use of Radio Frequency Identification (RFID) Tags in Bedside Monitoring of Endotracheal Tube Position. J Clin Monit Comput 21(3), 155-158 (2007)
14. Cousins, D.D.: Preventing Medication Errors. US Pharmacist, 70-75 (August 1995)

Affection-Based Visual Communication in the Mobile Environment

Hang-Bong Kang, Jung-Un Kim, Il-Whang Byun, Minjung Kim, Soo-Young Park,

Dept. of Digital Media, Catholic Univ. of Korea.
43-1 Yeoggok 2-dong, Wonmi-gu, Bucheon-si, Gyeonggi-do 420-743, Korea
hbkang@catholic.ac.kr

Abstract. In this paper, we propose a new visual communication method in the short messaging system. For example, text-based posts like Twitter in the limit of 140 characters are not efficient, though useful some times, to clearly express the author's message to his followers. To contend with this shortcoming, we suggest posting an appropriate image that can be placed with or without a text message. To generate an image with avatars, objects and backgrounds based on the text message, the keyword detector detects the key words and retrieves the thumbnail images that correspond with those keywords. After that, the author can select appropriate images for his background, characters and objects. To represent affection through images, we designed affection-based re-coloring method using interactive genetic algorithm. Our visual communication method is implemented on the i-phone and can post a tweet by using an image. The survey result shows that our method is more favorable in posting a message, than just a plain text message.

Keywords: short messaging system, visual communication, affection.

1 Introduction

Many social networking services such as Myspace, Facebook and Twitter are very popular to users in interacting with their friends or followers [1,2,3]. Recently, Twitter has gained popularity worldwide because it provides micro-blogging service or short message services that enable its users to exchange short messages as tweets in the PC as well as in the mobile environment. Tweets are text-based message up to 140 characters and delivered to the author's subscribers who are known as followers.

Even though text-based posts are useful to communicate with followers, it is not always easy to express one's own message in the limit of the 140 characters. For example, if one wants to post the tasty food or beautiful scenery to his followers, the text message is too short to represent his feelings. In addition, language differences can play a huge role in creating miscommunication when messages are delivered to foreign followers with different native languages [4]. Hence, it is desirable to use a visual communication method such as sketches and photos because visual images are an universal language.

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

For effective visual communication conveyed to one's followers, it is necessary to post pictures or images based on the author's own feelings. One way to represent affection in images is to transform colors of images, but it is not an easy task. It is necessary to learn the author's color perception and reflect it onto the image.

In this paper, we propose a new visual communication method in the mobile environment. We developed an image generation method that creates an image from the text message and transforms colors based on the author's affection. Our method is applied to Twitter. The rest of this paper is organized as follows. Section 2 explains visual communication in the social networking system. Section 3 presents our affection-based image re-colorring method. Section 4 discusses our experimental results.

2 Visual Communication in the Social Networking System

In the Social Networking System (SNS) like Twitter, many people post short messages regarding personal events. However, limited text space sometimes creates problems in effectively delivering one's story to others. In order to accurately describe events, specialized vocabulary is sometimes necessary. Using visual images is another effective way to describe events. Fig. 1 shows examples of the case of text message and the case of a visual image.

To represent short messages in SNS using images, we must analyze the text messages first. The topics of text messages are largely from personal events – A visit to a restaurant and its food, a visit to a friend's birthday party, or a visit to a theme park. In this message, we find four elements such as avatars, objects, backgrounds and the avatars' behavior. Thus, when we converted a short message to the images, we

(a) (b)

Fig. 1. (a) text-based message, (b) message using Image

generated the image consisting of four layers.

Currently, many smart phones provide touch interfaces for input. For visual communication, we need to consider two factors. First, we should use simple mechanism to generate an image on the smart phone for users as well as low power consumption. Secondly, we should take into account the resolution of the image. Usually, the image should fit the display size of the smart phone.

In this environment, our method for visual image generation is as follows: First, the user inputs his text message into the smart phone. Then, keywords will be detected in the text message. If a user selects one of them, relevant images will be retrieved from the DB. The image will be created from the layers of background, avatars, objects and the avatar's behavior. For the background, the user selects an image based on personal preference. On the selected background image, the user can place the avatar and objects by clicking keywords. At this stage, the user can move the position and size of the avatar and objects. Finally, the user selects behavior of the avatar by choosing from the templates provided. Fig. 2 shows the procedure of an image generation.

3 Affection-Based Visual Communication

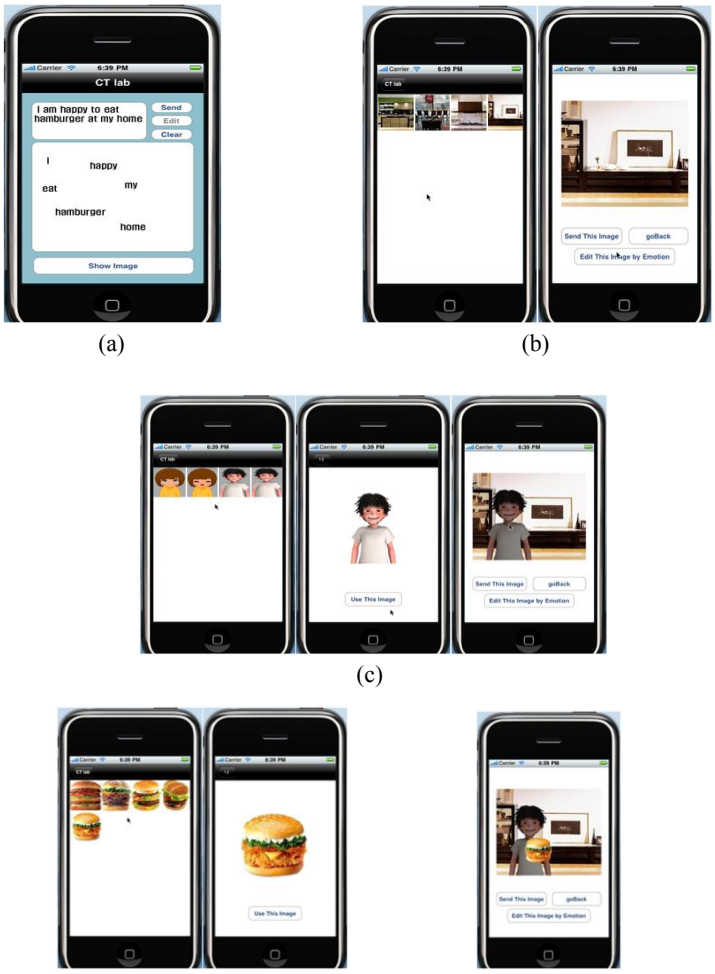
Whenever an author posts or sends a text message, his emotions are usually reflected in the message. To reflect these emotions onto an image for visual communication, we design a re-coloring method on the image based on the author's emotion. Usually, each person has different color preferences for each emotion [5,6,7]. Therefore, we first generate basic color templates for each emotion and then use a learning method to reflect the author's preferences. As a learning method, we use an interactive genetic algorithm.

To construct basic templates, 150 images are divided by 32 students into four basic emotions such as happiness, sadness, fear and anger. From those images, we intersect color histograms using 36 quantized HSV scheme and then finally construct basic color templates for four emotions. These templates are used to generate initial images for learning personal preferences. In this Section, we will discuss personalized color template using the interactive genetic algorithm.

The interactive genetic algorithm is a Genetic Algorithm where the evaluation part of it is subjectively handled by the user [8]. In fact, the user's preference cannot be numerically quantified because it depends on the perception of the user. So, the optimization is performed by human evaluation in the interactive genetic algorithm.

In our approach, we represent the chromosome of the color image by 7 bits. In the chromosome representation, 3 bits are assigned to represent 7 hue templates, 2 bits are assigned to represent 4 saturations and 4 values, respectively. Hence, 112 different representations are available. This is shown in Fig. 3.

The construction procedure of the personalized color templates with an interactive genetic algorithm is shown in Fig. 4. The process of each stage is as follows: Initially, the first generation of individuals is constructed from the basic templates. After that, the user evaluates the individuals. The evaluation scores are used as fitness values within the interactive genetic algorithm. From the evaluation results, the second



(a) (b)

(c)

(d) (e)

Fig. 2. Procedure of an image generation: (a) key word detection, (b) background image selection, (c) character selection, (d) object selection, (e) final image.

generation of individuals is made by the operations of genetic algorithm such as selection, crossover and mutation. This procedure is iterated and finally terminated by the user's decision when the user finds a desirable individual.

With the termination of the interactive genetic algorithm, we can extract chromosome information for H, S, and V. Finally, we can construct color templates which reflected the user's color preference for each emotion. Once the user's preference is learned in the training phase using interactive genetic algorithm, the constructed templates are used to generate personalized affection-based color transformed images.

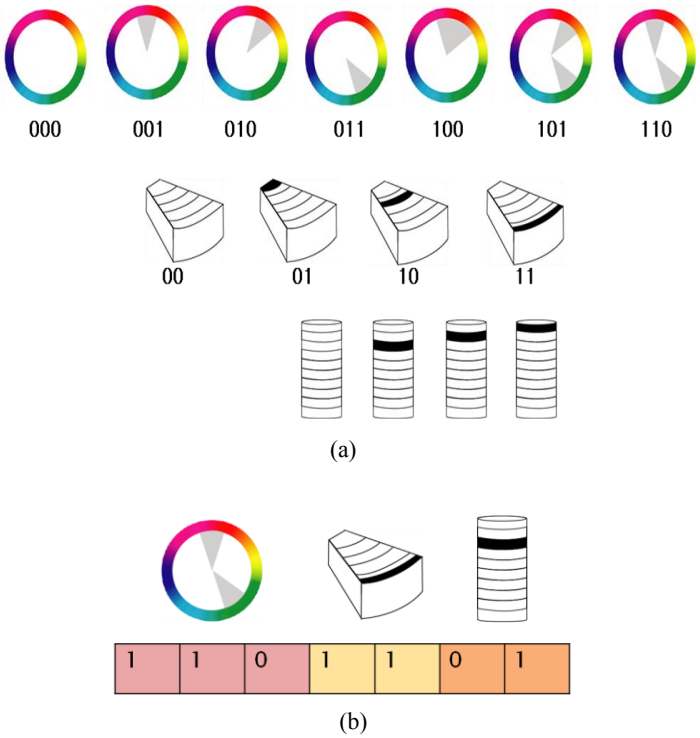


Fig. 3. Color Templates: (a) H, S, and V templates, (b) chromosome representation

4 Experimental Results

Our proposed method is implemented in the smart phone platform such as the i-phone. Fig. 5 shows an overview of our visual communication of tweets. First, the author types a text message into the phone. Then, the key word detector extracts key words from the message. After that, a group of thumbnail images are retrieved according to the keywords. The author can select the image he wants out of many background images, avatars and objects. In this phase, the author can modify the position and scale of the objects. Finally, an image instead of a text message is generated.

To reflect the affection of the author, a re-coloring method is implemented. The author can input his preference during the training phase. In the training phase, we can reflected the author's preferences using an interactive genetic algorithm. The author inputs his evaluation score from 0 to 5. If the author presses the keep button, the image is transferred to the next stage. After a few number of iterations, the author can stop the iteration and select the most favorable image. From that image, the

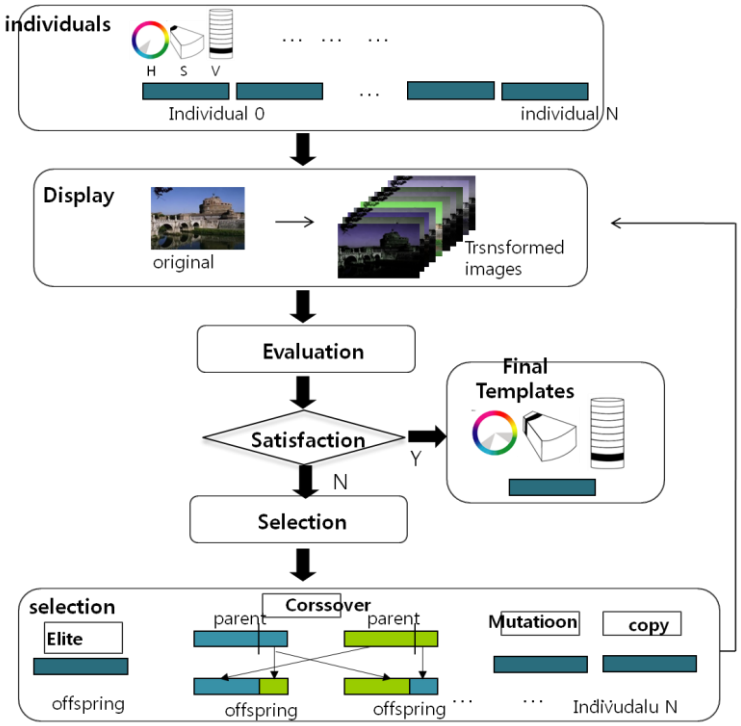


Fig. 4. Color preference learning using an interactive genetic algorithm

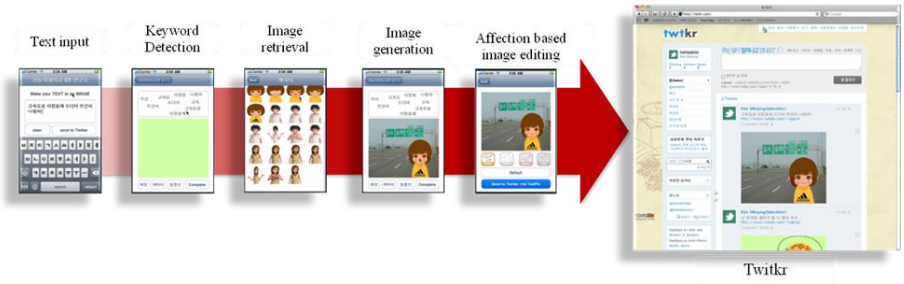


Fig. 5. Overview of our visual communication approach

modified color templates are constructed. Using these templates, the color transform is executed. Usually, the training is terminated after 3 or 4 iterations. The training phase is only executed by the user whenever he wants to reflect his color preferences. After the training, the author's preferred color image is generated based on his emotion. Fig. 6 shows an example of a re-colored image according to the mood of "happy".

Fig. 6. Examples of "Happy" mode.

To evaluate our method, we surveyed 15 students. 80% of students answered that the visual communication is more efficient than just using a plain text. 93% of students answered that the image generation process is easy. 60% of students thought that the training phase using an interactive genetic algorithm was easy. Even though the training phase requires the user's patience, 80% of students favored our affection-based re-colored image. Fig. 7 shows the survey results.

In order to post the visual image to Twitter, we first must send the generated image to the twitpic and then we will receive the URL information. After that, we can send the URL information to the http://www.twitkr.com. Finally, we can post a visual image with a text message like Fig. 8.

5 Conclusion

In this paper, we propose a new visual communication method in the short messaging system such as Twitter. Though transforming text to image is not an easy task, an image is more powerful communication tool than just words. To generate an image from a text message, the keyword detector detects the keywords and retrieves the thumbnail images according to those keywords. By selecting the images for background, avatar and objects, the final completed image is easily generated.

To reflect the author's emotion, we designed affection-based image re-coloring method using an interactive genetic algorithm. Based on the survey results from many subjects, we constructed basic color templates for each emotion. After that, the author of the image can evaluate the generated images from basic color templates. Based on the fitness score, the images are selected using tournament method and the off-springs are generated. After iterating this process, the author stops the process if he finds his favorable image. From that image, we construct color templates. These modified templates reflect the user's color preferences. Our experiments showed desirable results.

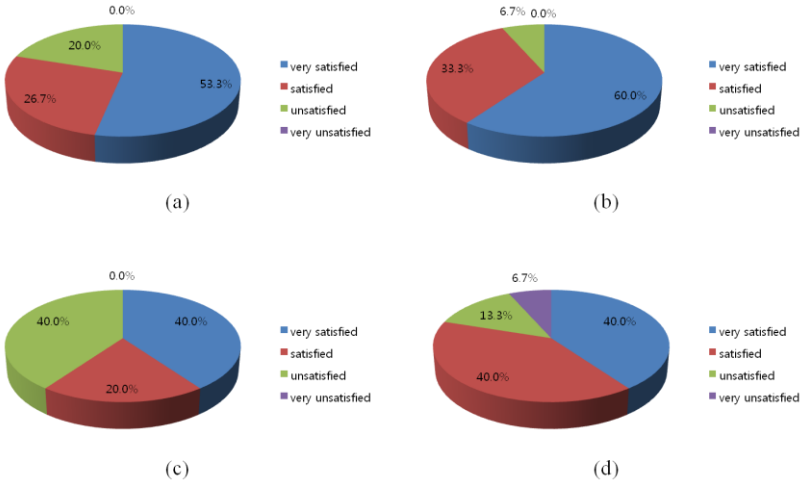


Fig. 7. Survey result of our method. (a) efficiency of visual communication, (b) easiness of image generation process, (c) easiness of training phase of interactive genetic algorithm, (d) preference of emotion-based re-colorring

Fig. 8. An example of visual tweet

In the future, it is desirable to implement sketch-based user interface to edit the images easily. In addition, cross-cultural research could shed the light on issues about how cultural differences vary in color-emotion associations.

Acknowledgments. This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Culture Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2009.

References

1. Comm, J.: Twitter Power 2.0: How to Dominate your Market One Tweet at a Time (2010)
2. Shih, C.: The Facebook Era: Tapping Online Social Networks to Build Better Products, Reach new Audiences, and Sell More Stuff (2009)
3. O'reily, T.: The Twitter Book (2009)
4. Gerarald, A., Goldstein, B.: Going Visual (2005)
5. Valdez, P., and Mehrabian, A.:Effects of color on emotions, Journal of Experimental Psychology: General, 394-409 (1994)
6. Picard, R.: Affective Computing. MIT Press (1997)
7. Kang, H.-B.: Affective Content Detection using HMMs, Proc. ACM Multimedia (2003)
8. Sugahara, M., Miki, M. and Hiroyasu, T.: Design of Japanese Kimono using Interactive Genetic Algorithm, Proc. IEEE Conf. Sys. Man and Cyber. (2008)

Development of an Intelligent e-Restaurant with Menu Recommendation for Customer-Centric Service

Tan-Hsu Tan[1,1], Ching-Su Chang[1], Yung-Fu Chen[2], Yung-Fa Huang[3] , Tsung-Yu Liu[4]

[1]Department of Electrical Engineering, National Taipei University of Technology, No.1, Sec. 3, Chung-hsiao E. Rd., Taipei, 10608,Taiwan, R.O.C.

[2]Department of Health Services Administration, China Medical University, No. 91, Hsuch-shih Rd., Taichung, 40402,Taiwan, R.O.C.

[3]Department of Information and Communication Engineering, Chaoyang University of Technology, No.168, Jifeng E. Rd., Wufeng Township, Taichung County, 41349, Taiwan, R.O.C.

{thtan, chingsu}@ntut.edu.tw, yungfu@mail.cmu.edu.tw, yfahuang@mail.cyut.edu.tw

[4]Department of Multimedia and Game Science, Lunghwa University of Science and Technology, No.300, Sec.1,Wanshou Rd., Guishan Shiang,Taoyuan County 33306, Taiwan, R.O.C.

prof_liu@pchome.com.tw

Abstract. Traditional restaurant service is generally passive: waiters must interact with customers directly before processing their orders. However, a high-quality customer-centered service system would actively identify customers and their favorite meals and expenditure records. To achieve this goal, this study integrates radio frequency identification (RFID), wireless local area network (WLAN), database technologies and a menu recommendation subsystem to develop an intelligent e-restaurant for customer-centric service. This system enables waiters to immediately identify customers via RFID-based membership cards and then to actively recommend the most appropriate menus for customers. Experimental results obtained from a case study conducted in a restaurant indicate that the proposed system has practical potential in providing customer-centric service.

Keywords: Radio Frequency Identification (RFID), Wireless Local Area Network (WLAN), Menu Recommendation Subsystem, Intelligent e-restaurant.

1 Introduction

Restaurant service such as making reservations, processing orders, and delivering meals generally require waiters to input customer information and then transmit the orders to kitchen for meal preparation. When the customer pays the bill, the amount due is calculated by the cashier. Although this procedure is simple, it may significantly increase the workload of waiters and even cause errors in meal ordering or in prioritizing customers, especially when the number of customers suddenly increases during busy hours, which can serious degrade the overall service quality. Therefore, using advanced

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

technologies to improve service quality has attracted much attention in recent years. For instance, the counter system of many fast food restaurants in Taiwan is equipped with a touch-screen, keypad or mouse control interface to enable cashiers to address customer needs. Such systems usually have common Point of Sale (POS) functions which allow waiters to use an optical scanner to directly read 2D barcodes for order details and billing. However, the POS system requires the waiter to determine customer needs and then enter the information. Therefore, service can only be provided passively. However, a high-quality service system should be customer-centered, i.e., it should immediately recognize the identities, favorite meals and expenditure records of customers so as to provide customer-centric services.

RFID has been identified as one of the ten greatest contributory technologies of the 21st century. It features more distant reading ability, larger memory capacities and reading ranges, and faster processing capability than the bar code system. RFID can also be used to identify objects or human beings. Due to its several advantages, RFID has been applied in many areas, such as supply chain management [1], telemedicine [2], manufacturing, inventory control [3], construction industry [4] and digital learning [5]. While RFID has successfully been employed in many areas, further exploration of its innovative applications is needed to enhance competitive advantage of enterprises and quality of life. For example, innovative applications of RFID are still rare in the restaurant industry. Recently, Ngai et al. [6] developed an RFID-based sushi management system in a conveyor-belt sushi restaurant to enhance competitive advantage. Their case study showed that RFID technology can help improve food safety, inventory control, service quality, operational efficiency and data visibility in sushi restaurants. Unfortunately, this system does not support customer-centered service because it can not actively identify customers.

In recent years, various product recommendation systems have been developed to enhance customer satisfaction and perceived value. Defined as a system which recommends an appropriate product or service after learning the customers' preferences and desires, recommendation systems are powerful tools that allow companies to present personalized offers to their customers. Extracting users' preferences through their buying behaviors and histories of purchased products is the most important element of such a system [7]. Wang et al. [8] proposed a recommendation system to avoid customer churn. In their study, different strategies can be made readily available that at once help maintain amiable customer relationships and suit new marketing conditions and circumstances.

To enhance customer service quality and improve competitiveness of restaurant industry, this study will integrate radio frequency identification (RFID), wireless local area network (WLAN), database technologies and a menu recommendation subsystem to implement an intelligent e-restaurant that enables waiters to immediately identify customers via their own RFID-based membership cards and then to actively recommend the most appropriate menus for customers. Customers can also use the RFID-based membership card to pay bills instead of using cash. Moreover, in order to enhance dining table service, the proposed system enables waiters to access customer information and make order by personal digital assistant (PDA). The PDA-based service unit enables instant transmission of customer

orders via WLAN to the kitchen for meal preparation. In addition, the expenditure information can be sent to the cashier for bill pre-processing. The restaurant managers can access the database to evaluate the business status anytime and make appropriate redeployments for food materials. Notably, all ordering and expenditure information is digitized for database storage, which allows restaurant owners to consider discounts or promotion to customers based on expenditure statistics. Customers can thus appreciate high-quality service, which in turn highly promotes enterprise image and increases business revenue for the restaurant. The rest of this paper is organized as follows. Section 2 describes the overview of the framework of the proposed system. Menu recommendation subsystem is demonstrated in Section 3. Evaluation results are given in Section 4. Finally, Section 5 presents the conclusions.

2 Proposed Intelligent e-Restaurant

Figure 1 shows an overview of the framework of our proposed intelligent e-restaurant for customer-centric service. This system provides online meal-ordering and reservation-making functions as well as personal meals recommendation service. The menus recommendation subsystem enables waiters to immediately identify customers via RFID-based membership cards and then to actively recommend the most appropriate menus for familiar customers according to their consuming records. Furthermore, the proposed system can recommend the most appropriate menus for new customers according to food material, price, and multi-criteria decision making analysis.

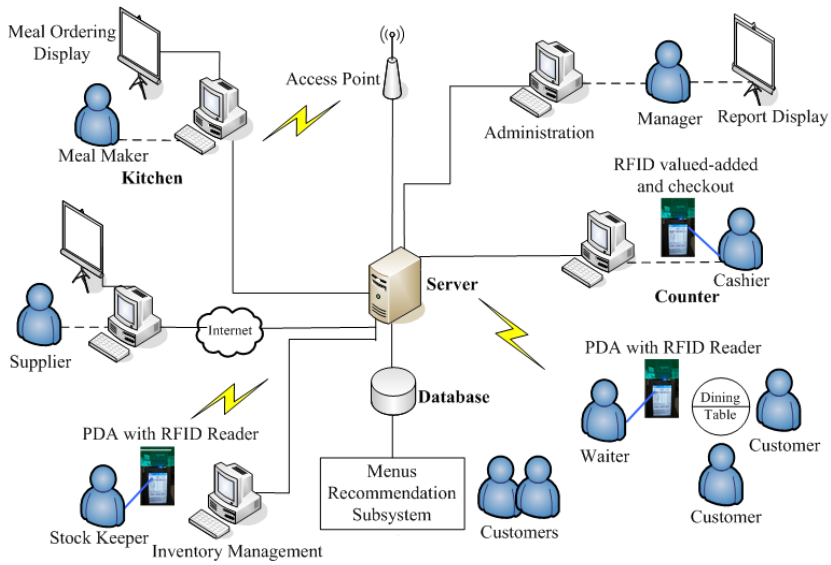


Fig. 1. Framework of intelligent e-restaurant for customer-centric service.

In this system, the waiter could use a PDA to make order for the customer at dining table and then wirelessly send the order to the kitchen server. The chefs could prepare the meal from the message shown on the order display system built in the kitchen. Furthermore, the system could be used to view statistics of the current inventory, sales records, staff information, and other information by the

restaurant manager. The stock-keeper could use a PDA to systematically record order information from the suppliers and monitor the available stocks in the shopping floors and the refrigerators. The system will send information and alert to the suppliers as inventory is lower than the stock level. After customer finishing the meal, the cashier could use an RFID-based PDA to identify the membership ID to check out the bill. In additional, the cashier could use the RFID-based PDA to perform the valued-added for customers.

3 Menu Recommendation Subsystem

3.1 Multi-criteria Decision Making Approach

The menu recommendation subsystem is developed based on multi-criteria decision making (MCDM) approach. MCDM [9-13] can be effectively utilized in the evaluation of alternatives (i.e., menus) because it evaluates items available for selection using multiple criteria (e.g., specifications). Thus, if a user directly states that he/she prefers one meal over the others, or if such information is indirectly obtained from the web usage behavior, it is reasonable to assume that the value of the preferred alternative is at least as great as that of the less preferred alternative. A well-known method for the evaluation of alternatives over multiple criteria is to use an additive form in which various values (scores or performances) of an alternative, measured with respect to each of the criteria, are added together to obtain an overall value of the alternative across multiple criteria, as defined in the following equation:

$$V(a_j) = \sum_{i=1}^{N} w_i v_i(a_j) \tag{1}$$

where V is the overall multiple criteria value, $0 \le V \le 1$; w_i is a scaling factor to represent the relative importance of the criterion; $v_i(a_j)$ is a single criterion value of alternative a_j with respect to criterion index i, $0 \le v_i(a_j) \le 1$. Accordingly, by considering two alternatives, a_k and a_m, the fact that a_k is preferred to a_m is denoted by

$$V(a_k) \ge V(a_m) \text{ or } \sum_{i=1}^{N} w_i v_i(a_k) \ge \sum_{i=1}^{N} w_i v_i(a_m) \tag{2}$$

Suppose two menus a_s and a_t are not included in the alternatives and need investigating to identify which menu the customer ranks more highly. To compare two alternatives under the feasible region of criteria weights, this work designs a PSO algorithm [14] to calculate the level of a_s over a_t on two extreme points, i.e., pessimistic and optimistic. The level of a_s over a_t in the interval $[\zeta_{\min}(a_s, a_t), \zeta_{\max}(a_s, a_t)]$ represents the range of possible differences of evaluation values between a_s and a_t.

In this work, the MCDM approach performs menu recommendation as follows:

Step 1: Extract the customer's implicit preference judgments from back-end database;
Step 2: Generate constraints set with implicit preference judgments;

Step 3: Use weighted fuzzy aggregation operators to infer the weight of each criterion for each constraint;

Step 4: Estimate the interval values $\{\zeta_{\min}(\cdot),\ \zeta_{\max}(\cdot)\}$ of the multi-criteria value function;

Step 5: Obtain the dominance values between criteria parameters;

Step 6: Transform the dominance values into the strength of the customer's preference for the provided menus;

Step 7: Use the OWA aggregation operator of the fuzzy linguistic quantifier to obtain the dominance degree (DD) between all parameters to be selected.

Step 8: A final decision can be made by the customer's DD such that the larger the DD of a menu, the better it is.

3.2 Generation of Ordered Weighted Averaging (OWA) Operator Weights with Fuzzy Linguistic Quantifier

Multi-criteria decision making (MCDM) problems are usually embedded with uncertainty. One of these uncertain parameters is the decision maker's degree of optimism, which has an important effect on decision outcomes. In this study, the fuzzy linguistic quantifiers [15] are used to obtain the assessments of this parameter from decision maker and then, because of its uncertainty it is assumed to have stochastic nature. Given a fuzzy linguistic quantifier Q, we can generate the Ordered Weighted Averaging (OWA) weights by $w_i = Q((i/N) - Q((i-1)/N))$, for i=1,$\cdots$, N ; then we can associate with this quantifier a degree of orness as below:

$$\text{orness}(Q)=\alpha = \sum_{i=1}^{N} \frac{N-i}{N-1}(Q(\frac{i}{N}) - Q(\frac{i-1}{N})).$$

In this study, the degree of domination of each alternative over the remaining ones is obtained by the means of aggregation with a Fuzzy OWA operator. The procedures of the algorithm are illustrated as follows.

Step1. Randomly generate N+1 nonnegative real number p_i with increasing order, i.e., $p_i - p_{i-1} > 0, (\ i = 1, 2, \cdots, N)$, and $p_0 = 0$;

Step2 Calculate $q_i = \sum_{k=1}^{i} p_k$, $s_i = q_i / q_N$ and $\alpha' = \sum_{k=1}^{N-1} \frac{s_k}{N-1}$, $(i = 0, 1, \cdots, N)$, respectively;

Step3. If $\alpha' \geq \alpha$, perform Step 4, otherwise perform Step 5.

Step4. Calculate $\alpha'' = \alpha / \alpha'$, $s_i' = s_i \times \alpha''$, $(i = 1, 2, \cdots, N-1)$ and $s_N' = s_N$.

Step5.

5.1. Let $s_i' = s_i + ir,\ (i = 1,\ 2, \cdots, N)$ and solve $\sum_{i=1}^{N-1} \frac{s_i' / s_N}{N-1} = \alpha$ for r,

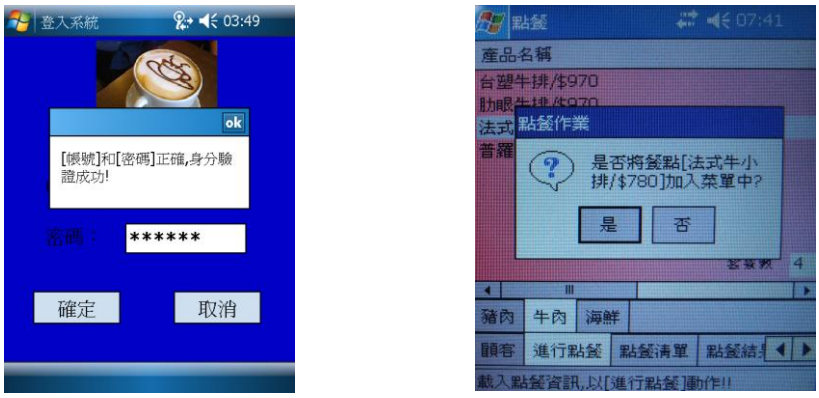
5.2. Calculate $w_i = (s_i' - s_{i-1}') / s_N'$,

5.3. end.

4 Evaluation results

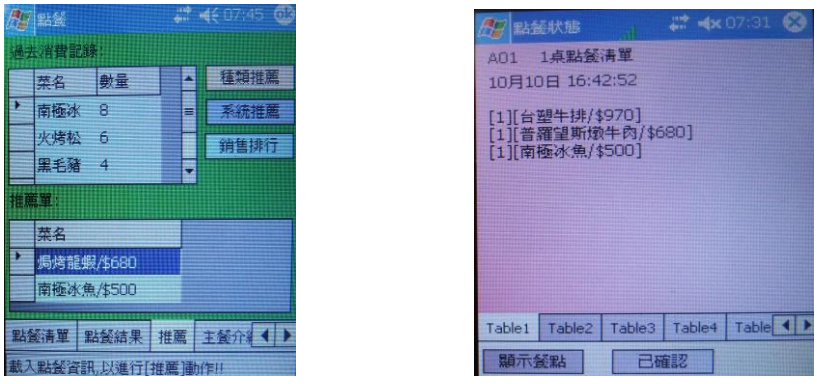
The user interface of the proposed system is built with Visual C# 2005 and eMbedded Visual C++, and the database is built on Microsoft SQL Server 2005 for server management and statistic reporting. Figure 2(a) and 2(b) show the system login interface on the client side and the ordering information of a customer, respectively. Figure 3(a) and 3(b) present the recommendation result from the customer's perspective and the menu ordering information displayed in kitchen side, respectively.

The menus recommendation subsystem enables waiters to immediately identify customers via RFID-based membership cards and then to actively recommend the most appropriate menus for customers according to their consumption records. For new customers, the service staff can provide recommendations based on food materials and meal popularity, and then create customers' preference to store in the back-end database; for long-time customers, service staff can use a MCDM approach to infer items preferred by customers or items close to those preferred items based on customers' preference data stored in the system. Therefore, this system is expected to help service providers increase their interactions with customers and provide fast and thoughtful services.



(a) (b)

Fig. 2. (a) Login interface and (b) ordering information of a customer.



(a) (b)

Fig. 3. (a) Recommendation result and (b) ordering information displayed in kitchen side.

The Technology Acceptance Model (TAM) [16] was employed to measure usefulness and ease of

use of the proposed system. The TAM is an information system (IS) that models how users come to accept and use a technology. The TAM posits that two particular beliefs, perceived ease of use and perceived usefulness, are of primary relevance. Perceived ease of use is the degree to which the prospective waiter perceives the IS as easy to use. Perceived usefulness is defined as the subjective belief that the use of a given information system improves waiter working efficiency by using the RFID-related consumer electronic products. The "attitude toward using" is a function of the perceived usefulness and perceived ease of use that directly influences actual usage behavior of customers.

A preliminary experiment has been conducted in a Taipei restaurant. A questionnaire was administered to fifteen waiters and forty-five customers to assess the ease of use (Part A), perceived usefulness (Part B) and attitudes toward the use of the proposed e-restaurant system (Part C). Responses were measured using a five-point Likert-scale from 1 (strong disagreement) to 5 (strong agreement). Table 1 lists the survey results of the questionnaire survey based on the 5-point scale. The questionnaire surveys are as follows:

1. Most waiters found the system interface to be user friendly (m=4.3).
2. Most waiters believed the proposed system was effective and easy to use for providing customer-centric service (m=4.6).
3. Most waiters believed the proposed system can accelerate the service process (m=4.6).
4. Most waiters thought the proposed system can improve working efficiency and service quality (m=4.7).
5. Most customers appreciated the real-time ordering and check out services provided by the proposed system (m=4.5).
6. Most customers appreciated the recommendation function (m=4.36) since it offers appropriate food choices to the customer.

Table 1. Statistical results of questionnaire.

Part	Item	Mean
A (ease of use) (Waiter)	A1. The interface is user-friendly.	4.3
	A2. The intelligent e-restaurant has sufficient functions and is easy to operate for customer-centric service.	4.6
B (usefulness) (Waiter)	B1. The intelligent e-restaurant improves the efficiency of the service process.	4.6
	B2. The intelligent e-restaurant can improve my working efficiency and service quality.	4.7
C (attitude toward using) (Customer)	C1. The intelligent e-restaurant can significantly reduce waiting time because it can provide real-time ordering and check out services.	4.5
	C2. The menus recommendation subsystem is helpful to make appropriate food choices.	4.36

5 Conclusions

This study constructed an intelligent e-restaurant system using RFID, WLAN, database technologies, and a menu recommendation subsystem to offer customer-centric service. A preliminary experiment has been conducted in a restaurant, and a questionnaire survey was administered to fifteen waiters and forty-five customers. The survey result is encouraging. In addition, we also conducted extensive interviews with restaurant owner and the results indicated that the proposed system is useful in reducing running cost, enhancing service quality as well as customer relationship. We will conduct a full scale experiment in the near future with more restaurants and improve system functions based on the experimental results and the participants' feedback to meet the requirement of practical application. Notably, the proposed system can easily be promoted to other areas, such as 3C product sales, livelihood-related industries, and department stores, via properly adjusting the e-restaurant framework, thus significantly enhancing the developments of RFID, information, and communication industries as well as national competitiveness.

References

1. M. Tajima, "Strategic Value of RFID in Supply Chain Management," Journal of Purchasing & Supply Management, vol. 13, no. 4, pp. 261--273 (2007)
2. Y. Xiao, X. Shen, B. Sun, and L. Cai, "Security and Privacy in RFID and Applications in Telemedicine," IEEE Communications Magazine, vol. 44, no. 4, pp. 64--72 (2006)
3. C.C. Chao, J.M. Yang, and W.Y. Jen, "Determining Technology Trends and Forecasts of RFID by a Historical Review and Bibliometric Analysis from 1991 to 2005," Technovation, vol. 27, no. 5, pp. 268--279 (2007)
4. K. Domdouzis, B. Kumar, and C. Anumba, "Radio-Frequency Identification (RFID) Applications: A Brief Introduction," Advanced Engineering Informatics, vol. 21, no. 4, pp. 350--355 (2007)
5. T.H. Tan, T.Y. Liu, and C.C. Chang, "Development and Evaluation of an RFID-based Ubiquitous Learning Environment for Outdoor Learning," Interactive Learning Environments, vol. 15, no. 3, pp. 253--269 (2007)
6. E.W.T. Ngai, F.F.C. Suk, and S.Y.Y. Lo, "Development of an RFID-based Sushi Management System: The Case of a Conveyor-belt Sushi Restaurant," International Journal of Production Economics, vol. 112, no. 2, pp. 630--645 (2008)
7. A. Albadvi, and M. Shahbazi, "A Hybrid Recommendation Technique Based on Product Category Attributes," Expert Systems with Applications, vol. 36, no. 9, pp. 11480--11488 (2009)
8. Y.F. Wang, D.A. Chiang, M.H. Hsu, C.J. Lin, and I.L. Lin, "A Recommender System to Avoid Customer Churn: A Case Study," Expert Systems with Applications, vol. 36, no 4, pp. 8071--8075 (2009)
9. D.H. Choi, and B.S Ahn, "Eliciting Customer Preferences for Products from Navigation Behavior on the Web: A Multicriteria Decision Approach with Implicit Feedback," IEEE Transactions on Systems, Man and Cybernetics, pp. 880--889 (2009)

10. P.Ya. Ekel, J.S.C. Martini, and R.M. Palhares, “Multicriteria Analysis in Decision Making Under Information Uncertainty,” Applied Mathematics and Computation, vol. 200, no. 2, pp. 501--516 (2008)
11. S.M. Chen, and C.H. Wang, “A Generalized Model for Prioritized Multicriteria Decision Making Systems,” Expert Systems with Applications, vol. 36, no. 3, pp. 4773--4783 (2009)
12. B. Malakooti, “Identifying Nondominated Alternatives with Partial Information for Multiple-objective Discrete and Linear Programming Problems,” IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, pp. 95--107 (1989)
13. B.S. Ahn, “Multiattribute Decision Aid with Extended ISMAUT,” IEEE Transactions on Systems, Man, and Cybernetics—PART A: Systems and Humans, vol. 36, no 3, pp. 507--520 (2006)
14. J. Y. Tsao, Estimation of Carrier Frequency Offset for Generalized OFDMA Uplink Systems Using PSO Algorithms. Master Thesis, Department of Electrical Engineering, National Taipei University of Technology, July 2009.
15. J. Malczewski, “Ordered Weighted Averaging with Fuzzy Quantifiers: GIS-based Multicriteria Evaluation for Land-use Suitability and Analysis,” International Journal of Applied Earth Observation and Geoinformation, vol. 8, pp. 270--277 (2006)
16. N. Park, R. Roman, S. Lee, and J.E. Chung, “User Acceptance of a Digital Library System in Developing Countries: An Application of the Technology Acceptance Model,” International Journal of Information Management, vol. 29, no. 3, pp. 196--209 (2009)

AR Registration for Video-Based Navigation

Yan Wang[1], Li Bai[1], Linlin Shen[2]

[1]School of Computer Science and IT, University of Nottingham, UK
[2]School of Computer Science & Software Engineering, Shenzhen University, China
{bai, yqw}@cs.nott.ac.uk, llshen@szu.edu.cn

Abstract. We have developed a novel video based personal navigation system for mobile automotive systems and handheld augmented reality applications, using a combination of computer vision and augmented reality techniques. With this type of navigation systems on PDAs or mobile phones, virtual road signs are superimposed onto the video of the real road scene. Such navigation systems allow the driver to travel to the destination by following the virtual signs in the video, offering a more intuitive and safer navigation solution. In this paper two methods for augmented reality registration of virtual signs onto the road in the video are described. The registration methods involve camera calibration and the pose estimation of either the camera or the reference object in the scene from visual information. Registration results on real road videos are presented.

Keywords: mobile automotive, video based navigation, vision & augmented reality.

1. Introduction

We have developed a personal navigation system for mobile automotive systems and handheld augmented reality applications, using a combination of computer vision and augmented reality (AR) techniques. AR allows computer-generated graphics to be superimposed onto the image of the real scene. The graphics can be a virtual car for the driver to follow, or a virtual arrow superimposed onto the road scene to indicate travel directions. One major advantage gained from video based navigation is that it provides 'what you see is what you get' navigation, which retains the driver's awareness of road conditions and potential hazards, thus reducing the chances of road accidents. To correctly superimpose/register a virtual object onto the video, a reference object in the scene needs to be identified and the position and pose of the reference object with respect to the camera need to be estimated. In other words, AR registration is treated as a camera/object pose estimation problem. To do this, some known 3D geometric features must be matched to their image counterparts. Common features include corners/points [1][2], lines [3][4] and regions [5]. Camera/object pose estimation can be solved using either linear [6][7] or nonlinear techniques [8][9]. The

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

latter often involve nonlinear optimizations to minimize matching errors based on an initial solution by the former.

In this paper, we present two AR registration methods of virtual arrows on the road in the video using visually detected and tracked road features. We use the robust road tracking methods described in [10][11]. The road is represented as a hyperbola and tracked using the particle filter tracker. The AR registration methods are applicable to different road models such as clothoids [12], parabola [13], splines [14], and connected arcs [15], as long as these models support the parallel road boundary assumption. The tasks also include estimating the transformation matrix from the object reference frame to the camera reference frame. The paper is organized as follows. In Section 2 AR registration methods are presented. Section 3 presents experiment results, and Section 4 concludes the paper.

2. Registration

Four coordinate systems are involved in the registration: the object coordinate system (OCS), the road (i.e., the reference object) coordinate system (RCS), the camera coordinate system (CCS) and the image coordinate system (ICS), shown in Figure 1.

The registration of virtual arrows can be achieved in two different ways: projection and back-projection. The first approach transforms virtual arrows from OCS to CCS and projects them into the image space. The second approach back-projects image features of the reference object from the image to RCS. The transformation from OCS to CCS can be directly recovered from the back-projection process. For clarity, some symbols are defined in Table 1, where l_1, l_2 and $\boldsymbol{V}$ in Table 1 are assumed to be known from road detection. The road width w is assumed to be known.

Table 1. Symbols in Figure 1.

- $\boldsymbol{O}_c$, x_c, y_c, z_c: origin and axes of CCS.
- o, u, v: origin and axes of ICS.
- $\boldsymbol{A}$: orthogonal projection of $\boldsymbol{O}_c$ onto the ground plane.
- $\boldsymbol{O}_r$, x_r, y_r, z_r: origin and axes of RCS. $\boldsymbol{O}_r$ is the intersection of road midline with the projection of x_c onto the ground plane. y_r is parallel to the road tangent at $\boldsymbol{O}_r$. z_r is perpendicular to the ground plane.
- $\boldsymbol{B}$, $\boldsymbol{C}$: intersections of road boundaries with projection of x_c on the ground plane.
- $\boldsymbol{D}$ (u_D,v_D): intersection of v axis with vanishing line of the ground plane, represented by a unit vector $\boldsymbol{m}_D$.
- $\boldsymbol{n}_B$ and $\boldsymbol{n}_C$: parallel road tangents at $\boldsymbol{B}$ and $\boldsymbol{C}$.
- $\boldsymbol{E}$, $\boldsymbol{F}$: intersections of a line parallel to x_r through $\boldsymbol{A}$ with $\boldsymbol{n}_B$ and $\boldsymbol{n}_C$. $\boldsymbol{n}_B, \boldsymbol{n}_C \perp \overrightarrow{\boldsymbol{EF}}$. $\|\boldsymbol{EF}\|$ is road width w.
- $\boldsymbol{\zeta}$: unit vector from $\boldsymbol{O}_c$ to $\boldsymbol{E}$. $\boldsymbol{\zeta} = [\boldsymbol{n}_l \times \boldsymbol{m}_V]$.
- $\boldsymbol{\xi}$: unit vector from $\boldsymbol{O}_c$ to $\boldsymbol{F}$. $\boldsymbol{\xi} = [\boldsymbol{m}_V \times \boldsymbol{n}_r]$.
- $\boldsymbol{v}$: orthogonal projection of z_r on the ground plane. It forms angle ψ with z_r and angle α with

y_r respectively.

- $\boldsymbol{\mu}$: unit vector from $\boldsymbol{O}_c$ to $\boldsymbol{A}$, parallel to z_r:
- $\boldsymbol{\mu} = [\boldsymbol{m}_V \times \boldsymbol{v}] = \left(0, 1/\sqrt{1+v_V^2}, -v_V/\sqrt{1+v_V^2}\right)^T$.
- $\boldsymbol{P}$ $(x_P, y_P, 0)$, x_o, y_o, z_o: origin and axes of OCS.
- $\boldsymbol{n}_{xo}$, $\boldsymbol{n}_P$, $\boldsymbol{n}_{zo}$: unit vectors on x_o, y_o and z_o. $\boldsymbol{n}_{zo}=(0,0,1)^T$ and $\boldsymbol{n}_{xo} = \boldsymbol{n}_P \times \boldsymbol{n}_{zo}$.
- φ: yaw angle of the camera between the projection of z_c on the ground plane and the road direction.
- ψ: inclination angle between z_c and its orthogonal projection on the ground plane.
- l_1, l_2: road image tangents at $\boldsymbol{B}$ and $\boldsymbol{C}$, represented by unit vectors $\boldsymbol{n}_l$ and $\boldsymbol{n}_r$.
- $\boldsymbol{V}$ (u_V, v_V): intersection of l_1 and l_2, located on the vanishing line of the ground plane, represented by a unit vector $\boldsymbol{m}_V$. $\boldsymbol{m}_V = [\boldsymbol{n}_l \times \boldsymbol{n}_r]$. $\boldsymbol{m}_V = \boldsymbol{n}_B = \boldsymbol{n}_C$.

2.1. Registration from Projection

The projection method contains the following steps: (1) estimate the transformation from OCS to RCS; (2) estimate the transformation from RCS to CCS; (3) project the result from CCS to ICS. Figure 1 shows the geometry related to the transformations. The road has been detected using hyperbolas [12].

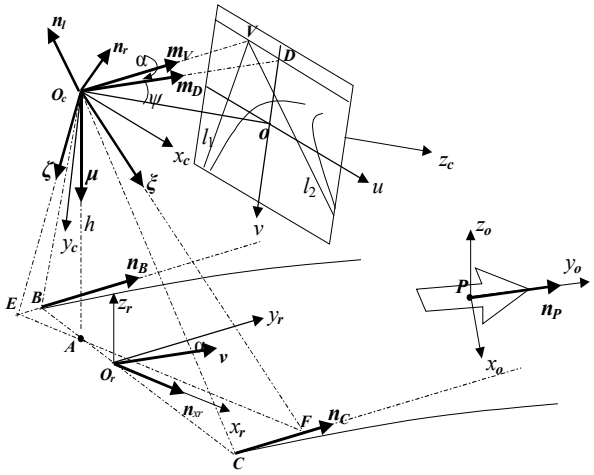


Fig. 1. Registration from projection.

We start from the transformation from OCS to RCS defined as a 4×4 matrix:

$$ {}^r\boldsymbol{M}_o = \begin{bmatrix} {}^r\boldsymbol{R}_o & {}^r\boldsymbol{t}_o \\ 0 & 1 \end{bmatrix}. \tag{2.1} $$

where ${}^r\boldsymbol{R}_o$ is a 3×3 rotation matrix and ${}^r\boldsymbol{t}_o$ a 3×1 translation vector:

$$ {}^{r}\boldsymbol{R}_o = \left[\boldsymbol{n}_{xo} \quad \boldsymbol{n}_P \quad \boldsymbol{n}_{zo}\right], {}^{r}\boldsymbol{t}_o = \overrightarrow{\boldsymbol{O}_r\boldsymbol{P}} = \left(x_P, y_P, 0\right)^T \ . \tag{2.2} $$

We then calculate the transformation matrix ${}^{c}\boldsymbol{M}_r$ from RCS to CCS. It has the same form as (2.1) with ${}^{c}\boldsymbol{R}_r$, the rotation matrix and ${}^{c}\boldsymbol{t}_r$, the translation vector. ${}^{c}\boldsymbol{R}_r$ can be parameterized as angles α and ψ.

$$ \begin{aligned} &{}^{c}\boldsymbol{R}_r = \boldsymbol{R}_{xr}(\pi/2)\boldsymbol{R}_{xr}(\psi)\boldsymbol{R}_{zr}(\alpha), \\ &\boldsymbol{R}_{xr}(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}, \boldsymbol{R}_{zr}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{2.3} $$

ψ and α can be determined as soon as $\boldsymbol{v}$ is solved. We have:

$$ \begin{aligned} &\boldsymbol{v} = \left(u_D, v_D, 1\right)^T \Big/ \sqrt{u_D^2 + v_D^2 + 1} = \left(0, v_V, 1\right)^T \Big/ \sqrt{v_V^2 + 1}, \\ &\left|\psi\right| = \left|\arccos\left(\boldsymbol{v} \cdot \boldsymbol{n}_{zc}\right)\right|, \left|\alpha\right| = \left|\arccos\left(\boldsymbol{v} \cdot \boldsymbol{m}_V\right)\right| \end{aligned} \tag{2.4} $$

$\boldsymbol{n}_{zc}$ is a unit vector $(0,0,1)^{\mathrm{T}}$ along z_c. The sign of ψ is inverse to the sign of v_D. The sign of α is identical to the sign of u_V.

We now solve for ${}^{c}\boldsymbol{t}_r$. It can be calculated as:

$$ {}^{c}\boldsymbol{t}_r = \overrightarrow{\boldsymbol{O}_c\boldsymbol{O}_r} = \overrightarrow{\boldsymbol{O}_c\boldsymbol{A}} - \overrightarrow{\boldsymbol{O}_r\boldsymbol{A}} = h\boldsymbol{\mu} - \overrightarrow{\boldsymbol{O}_r\boldsymbol{A}} \ . \tag{2.5} $$

Since $\boldsymbol{m}_V$ is parallel to y_r and $\boldsymbol{\mu}$ is parallel to z_r, one may work out a unit vector $\boldsymbol{n}_{xr}$ along x_r as $\boldsymbol{n}_{xr} = \left[\boldsymbol{\mu} \times \boldsymbol{m}_V\right]$. Consider two triangles $\boldsymbol{O}_c\boldsymbol{E}\boldsymbol{A}$ and $\boldsymbol{O}_c\boldsymbol{A}\boldsymbol{F}$, it can be seen that:

$$ \begin{aligned} &\frac{\left\|\overrightarrow{\boldsymbol{E}\boldsymbol{A}}\right\|}{h} = \left|\frac{\boldsymbol{\zeta} \cdot \boldsymbol{n}_{xr}}{\boldsymbol{\zeta} \cdot \boldsymbol{\mu}}\right|, \frac{\left\|\overrightarrow{\boldsymbol{A}\boldsymbol{F}}\right\|}{h} = \left|\frac{\boldsymbol{\xi} \cdot \boldsymbol{n}_{xr}}{\boldsymbol{\xi} \cdot \boldsymbol{\mu}}\right|, \\ &\frac{\left\|\overrightarrow{\boldsymbol{E}\boldsymbol{A}}\right\|}{h} + \frac{\left\|\overrightarrow{\boldsymbol{A}\boldsymbol{F}}\right\|}{h} = \frac{\left\|\overrightarrow{\boldsymbol{E}\boldsymbol{F}}\right\|}{h} \Rightarrow h = w\left(\frac{\boldsymbol{\xi} \cdot \boldsymbol{n}_{xr}}{\boldsymbol{\xi} \cdot \boldsymbol{\mu}} - \frac{\boldsymbol{\zeta} \cdot \boldsymbol{n}_{xr}}{\boldsymbol{\zeta} \cdot \boldsymbol{\mu}}\right)^{-1} \end{aligned} \tag{2.6} $$

We now need to determine lateral position $\boldsymbol{O}_r\boldsymbol{A}$:

$$ \begin{aligned} &\overrightarrow{\boldsymbol{O}_r\boldsymbol{A}} = \overrightarrow{\boldsymbol{O}_r\boldsymbol{B}} + \overrightarrow{\boldsymbol{B}\boldsymbol{A}} = \left(-\frac{w}{2\cos\alpha} + \frac{-\boldsymbol{\zeta} \cdot \boldsymbol{n}_{xr}}{\boldsymbol{\zeta} \cdot \boldsymbol{\mu}} \frac{h}{\cos\alpha}\right)\boldsymbol{n}_{xc} \\ &= \left[\frac{\boldsymbol{\zeta} \cdot \boldsymbol{n}_{xr}}{\boldsymbol{\zeta} \cdot \boldsymbol{\mu}} - \frac{\boldsymbol{\xi} \cdot \boldsymbol{n}_{xr}}{\boldsymbol{\xi} \cdot \boldsymbol{\mu}}\right]\frac{h}{2\cos\alpha}\boldsymbol{n}_{xc} = \frac{-w}{2\boldsymbol{m}_V \cdot \boldsymbol{v}}\boldsymbol{n}_{xc} \end{aligned} \tag{2.7} $$

where $\boldsymbol{n}_{xc}$ is a unit vector $(1,0,0)^{\mathrm{T}}$ along x_c. (2.5) thus becomes:

$$ {}^{c}\boldsymbol{t}_r = w\left(\frac{\boldsymbol{\xi} \cdot \boldsymbol{n}_{xr}}{\boldsymbol{\xi} \cdot \boldsymbol{\mu}} - \frac{\boldsymbol{\zeta} \cdot \boldsymbol{n}_{xr}}{\boldsymbol{\zeta} \cdot \boldsymbol{\mu}}\right)^{-1}\boldsymbol{\mu} + \frac{w}{2\boldsymbol{m}_V \cdot \boldsymbol{v}}\boldsymbol{n}_{xc} \ . \tag{2.8} $$

Note that ${}^{c}\boldsymbol{R}_r$ can be alternatively constructed by three column vectors:

$$ {}^{c}\boldsymbol{R}_r = \left[\boldsymbol{n}_{xr} \quad \boldsymbol{m}_V \quad -\boldsymbol{\mu}\right] \ . \tag{2.9} $$

Finally, we concatenate ${}^{c}\boldsymbol{M}_r$ and ${}^{r}\boldsymbol{M}_o$:

$$ {}^{c}\boldsymbol{M}_o = {}^{c}\boldsymbol{M}_r \, {}^{r}\boldsymbol{M}_o \ . \tag{2.10} $$

2.2. Registration from Back-projection

The back projection method derives the transformation from a back projection process without explicitly estimating the camera's pose. Figure 2 illustrates the process. For clarity, some symbols are defined in Table 2.

Table 2. Symbols in Figure 2.

- $\boldsymbol{O}_c$, x_c, y_c, z_c: origin and axes of CCS.
- $\boldsymbol{M}$ (u_M,v_M): selected image location for registration, represented by a unit vector $\boldsymbol{m}_M$.
- $\boldsymbol{A}$, $\boldsymbol{B}$: road boundary points with the same v coordinate as $\boldsymbol{M}$.
- l_A, l_B: road tangents at $\boldsymbol{A}$ and $\boldsymbol{B}$, represented by unit vector $\boldsymbol{n}_A$ and $\boldsymbol{n}_B$.
- $\boldsymbol{P}$, $\boldsymbol{\zeta}$, $\boldsymbol{\xi}$, $\boldsymbol{\omega}$: origin and axes of OCS. $\boldsymbol{P}$ is the back-projection of $\boldsymbol{M}$ on the ground plane. $\boldsymbol{\xi}$ is along the road tangential direction at $\boldsymbol{P}$. $\boldsymbol{\omega}$ is perpendicular to the ground plane:
- $\boldsymbol{\omega}=[0,-1,v_H]^T$. $\boldsymbol{\xi} = [\boldsymbol{n}_A \times \boldsymbol{n}_B]$. $\boldsymbol{\zeta}=\boldsymbol{\xi}\times\boldsymbol{\omega}$.
- $\boldsymbol{T}$: image point represented by $\boldsymbol{\zeta}$. It is on the vanishing line since $\boldsymbol{\zeta}$ is parallel to the ground plane.
- l_T: image line through $\boldsymbol{T}$ and $\boldsymbol{M}$, represented by $\boldsymbol{n}_T$:
- $\boldsymbol{n}_T = [\boldsymbol{\zeta} \times \boldsymbol{m}_M]$
- $\boldsymbol{V}$: intersection of l_A with l_B, represented by $\boldsymbol{m}_V$: $\boldsymbol{m}_V = \boldsymbol{\xi}$.
- $\boldsymbol{C}$, $\boldsymbol{D}$: back projections of $\boldsymbol{A}$ and $\boldsymbol{B}$ on the road.
- $\boldsymbol{n}_C$ and $\boldsymbol{n}_D$: road tangents at $\boldsymbol{C}$ and $\boldsymbol{D}$: $\boldsymbol{n}_C = \boldsymbol{n}_D = \boldsymbol{\xi}$.
- $\boldsymbol{E}$, $\boldsymbol{F}$: intersections of the line parallel to $\boldsymbol{\zeta}$ through $\boldsymbol{P}$ with $\boldsymbol{n}_C$ and $\boldsymbol{n}_D$, respectively. $\|\boldsymbol{EF}\|=w$, the road width. $\boldsymbol{G}$, $\boldsymbol{H}$: image projections of $\boldsymbol{E}$ and $\boldsymbol{F}$, which are also the intersections of l_T *with* l_A and l_B, represented by unit vectors $\boldsymbol{m}_G$, $\boldsymbol{m}_H$:

$$\boldsymbol{m}_G = [\boldsymbol{n}_A \times \boldsymbol{n}_{TM}] = [\boldsymbol{n}_A \times (\boldsymbol{\zeta} \times \boldsymbol{m}_M)] \,. \tag{2.11}$$

$$\boldsymbol{m}_H = [\boldsymbol{n}_B \times \boldsymbol{n}_{TM}] = [\boldsymbol{n}_B \times (\boldsymbol{\zeta} \times \boldsymbol{m}_M)] \,. \tag{2.12}$$

$\boldsymbol{A}$, $\boldsymbol{B}$, l_A and l_B in Table 2 are known from road detection. The transformation ${}^c\boldsymbol{M}_o$ is defined as:

$${}^r\boldsymbol{M}_o = \begin{bmatrix} \zeta & \xi & \omega & \overrightarrow{O_cP} \\ 0 & 0 & 0 & 1 \end{bmatrix} \,. \tag{2.13}$$

We now only need to calculate $\boldsymbol{O}_c\boldsymbol{P}$:

$$\overrightarrow{\boldsymbol{O}_c\boldsymbol{P}} = -h\frac{(\boldsymbol{\omega}\cdot\boldsymbol{m}_H)\boldsymbol{m}_G + (\boldsymbol{\omega}\cdot\boldsymbol{m}_G)\boldsymbol{m}_H}{(\boldsymbol{\omega}\cdot\boldsymbol{m}_G)(\boldsymbol{\omega}\cdot\boldsymbol{m}_H)} \,. \tag{2.14}$$

$\boldsymbol{m}_G$ and $\boldsymbol{m}_H$ are calculated using (2.11) and (2.12). Since

$$w = \left|\overrightarrow{\boldsymbol{EF}}\right| = \left|\overrightarrow{\boldsymbol{O}_c\boldsymbol{F}} - \overrightarrow{\boldsymbol{O}_c\boldsymbol{E}}\right| = \frac{h\left|\boldsymbol{\omega}\times(\boldsymbol{m}_G\times\boldsymbol{m}_H)\right|}{(\boldsymbol{\omega}\cdot\boldsymbol{m}_G)(\boldsymbol{\omega}\cdot\boldsymbol{m}_H)} \,. \tag{2.15}$$

Canceling h by dividing (2.14) over (2.15) yields:

$$\frac{\overrightarrow{\boldsymbol{O}_c\boldsymbol{P}}}{w} = -\frac{(\boldsymbol{\omega}\cdot\boldsymbol{m}_H)\boldsymbol{m}_G + (\boldsymbol{\omega}\cdot\boldsymbol{m}_G)\boldsymbol{m}_H}{2\left|\boldsymbol{\omega}\times(\boldsymbol{m}_G\times\boldsymbol{m}_H)\right|} \,. \tag{2.16}$$

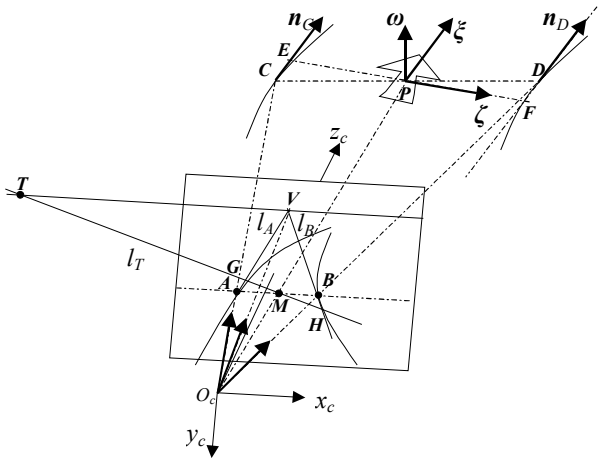


Fig. 2. Registration from back-projection.

3. Experiments

The system is implemented on a laptop with a 1GHz Intel® Pentium M processor. The other devices used include a GPS receiver to provide global positions at 1 Hz and a color camera to output a real-time video stream. The video frame size is 720×480 pixels. The camera was stabilized in a car facing towards the road in front. The two devices were connected to the laptop via USB and firewire respectively. The computer program contains several components: road detection and tracking, GIS data processing, and AR registration. AR registration is implemented using C/C++ and an augmented reality library ARToolkit based on OpenGL. At present the system works at a rate of 15 fps (video frames per second) whist registration only takes on average 2ms. The two registration methods can be easily implemented on mobile phones as the calculation is straightforward and does not require a high computation capacity.

The first two rows in Figure 3 show road-tracking and registration results on a set of video frames. As the vehicle was driven on a smooth lane, a set of virtual arrows (in yellow) moves forward along the midline of the current lane. The next four rows show navigation through road singularities such as road junctions (3rd and 4th row) and roundabouts (5th and 6th row). When the vehicle approaches a road junction, arrows are overlaid indicating the 'turn left'. After the vehicle turned left, arrows are overlaid indicating 'straight ahead'. At a roundabout, arrows are overlaid indicating 'turn left'.

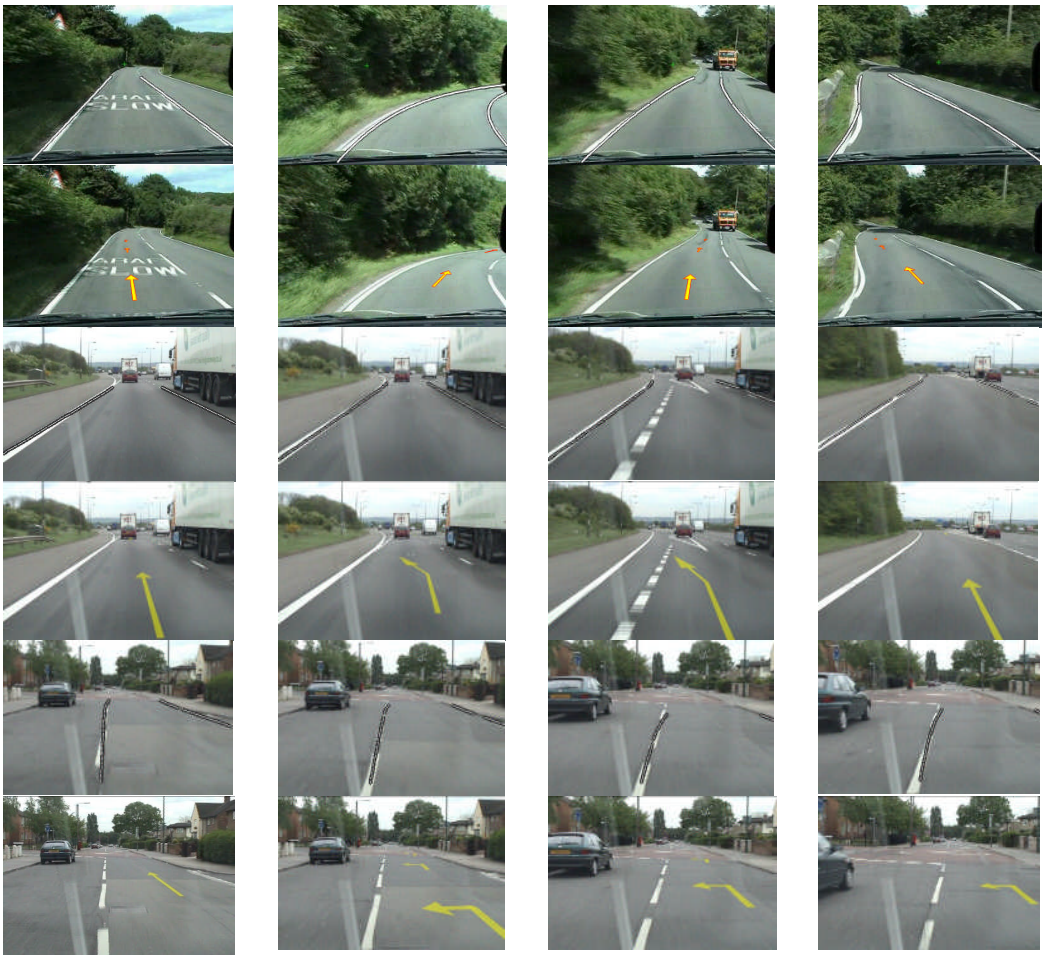


Fig. 3. Example road tracking and AR registration.

4. Conclusion

We have presented two AR registration methods for a novel video based personal navigation system for mobile automotive applications. It has been demonstrated that visual road detection and tracking can help AR registration of virtual arrows to navigate the driver through different road conditions. The registration methods can be easily implemented on handheld devices such as PDAs and mobile phones with a built-in camera. The methods presented here can be generalised for camera pose estimation for other mobile applications.

References

1. Nister, D.: An Efficient Solution to the Five-Point Relative Pose Problem, IEEE Transactions on Pattern Recognition and Machine Intelligence, 26(6):756-770 (2004)
2. Segvic, S., Remazeilles, A., Diosi, A., Chaumette, F. : A Mapping and Localization Framework for Scalable Appearance-Based Navigation, Computer Vision and Image Understanding, 113:172-187 (2009)
3. Christy, S., Horaud, R.: Iterative Pose Computation from Line Correspondences, Journal of Computer Vision and Image Understanding, 73(1):137-144 (1999)
4. Klein, G., Drummond, T.: Tightly Integrated Sensor Fusion for Robust Visual Tracking, Image and Vision Computing, 22:769-776 (2004)
5. Ferrari, V., Tuytelaars, T., Gool, L.: Markerless Augmented Reality with A Real-Time Affine Region Tracker, The Inter. Symp. on Augmented Reality, p.87 (2001)
6. Ji, Q., Costa, M.S., Haralick, R.M., Shapiro, L.G.: An Integrated Linear Technique for Pose Estimation from Different Geometric Features, International Journal of Pattern Recognition and Artificial Intelligence, 13(5):705-733 (1999)
7. Ansar, A., Daniilidis, K.: Linear Pose Estimation from Points or Lines, IEEE Transactions on Pattern Analysis and machine Intelligence, 25(5):578-589 (2003)
8. Drummond, T., Cipolla, R.: Real-Time Tracking of Complex Structures with On-Line Camera Calibration, BMVC (1999)
9. Comport, A., Marchand, E., Chaumette, F.: A Real-Time Tracker for Markerless Augmented Reality, Proceedings of Int. Symp. on Mixed and Augmented Reality, pp.36-45 (2003)
10. Bai, L., Wang, Y.: An Extended Hyperbola Model for Road Tracking, Journal of Knowledge-based Systems, 21(3):265-272 (2008)
11. Wang, Y., Bai, L., Fairhurst, M.: Robust Road Detection and Tracking using Condensation, IEEE Transactions on Intelligent Transportation Systems, 9(4):570-579 (2008)
12. Eidehall, A., Gustafsson, F.: Obtaining Reference Road Geometry Parameters from Recorded Sensor Data, IEEE Intelligent Vehicles Symposium (2006)
13. McCall, J.C., Trivedi, M.M.: Video Based Lane Estimation and Tracking for Driver Assistance: Survey, System and Evaluation, IEEE Transactions on Intelligent Transportation Systems, 2006
14. Wang, Y., Teoh, E.K., Shen, D.: Lane Detection and Tracking Using B-Snake, Image and Vision Computing, 22:269-280 (2004)
15. Wang, Y., Bai, L.: Fusing Image, GPS and GIS for Road Tracking Using Multiple Condensation Particle Filters, IEEE Intelligent Vehicles Symposium (2008)

An Efficient Seal Detection Algorithm

Zhimao Yao, Yonghong Song, Yuanlin Zhang, Yuehu Liu

Institute of Artificial Intelligence and Robotics,
Xi'an Jiaotong University, Xi'an, P.R.China, 710049
zmyao@aiar.xjtu.edu.cn, songyh@mail.xjtu.edu.cn

Abstract. This paper presents a novel seal detection algorithm which utilizes statistic shape features and can be efficiently used in the application of the mobile device. To extract different types of seals from the document image, the candidate regions of seal are located by using the foreground and background connected components analysis, and for rectangle seal, we improve the classic Hough transform method using the run-length histogram features to detect line segments and build a straight line relationship matrix, for circle and ellipse seal, its center is fixed efficiently using run-length and symmetry, and circle and ellipse seals can be detected by RANSAC fitting. The experimental results show that proposed algorithm can obtain high average-recall and precision-rates simultaneously.

Keywords: seal detection, run-length histogram, Hough transform

1 Introduction and Related Work

Documents which contain seals are very common in the current life, e.g. Seal detection should be used in electrical and postal business with the mobile device. Seals could supply us with a lot of useful information. Thus in this paper we propose a high efficient seal detection algorithm to automatically detect and segment seals in document images.

Seals generally consist in geometric shapes, such as circular or rectangular shapes. Our work focused on detecting some specific shapes of seal. Actually, early researches in seal detection mainly depend on color information [7-10]. However, the color information would often be lost during the process of document copying and transmission. In many cases, it is not reliable to use color information for seal detection. Traditional algorithms for shape detection can be divided into two main groups. The first group includes many variants of Hough transform method [1-3] and RANSAC method [4]. These methods are based on the voting mechanism, and commonly they are computationally expensive due to the high-dimension of the shape's parameter space. The other group of methods includes the shape fitting algorithm [5] and the genetic algorithm [6]. These methods are more efficient. However, they are usually sensitive to the noise. Most recently, researchers proposed

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

a robust seal detection framework to detect circular and elliptical seals in paper [11]. The method shows a good performance on processing time, but only the rough position of seals could be predicted.

In this paper, we proposed an efficient algorithm to detect and extract seals in document images based on the shape features. It mainly includes two stages, connected components analysis and shape analysis of seals. Seals in the document image are located by using connected components analysis first. Then these locations are further confirmed by the shape analysis. Based on the run-length histogram, an improved Hough transform is used to detect the straight lines. In this paper, three types of seals are analyzed and tested, i.e., circular, rectangular and elliptical seals, which are the most common ones in daily life. A flow chart of the proposed seal detection algorithm is shown in Fig. 1.

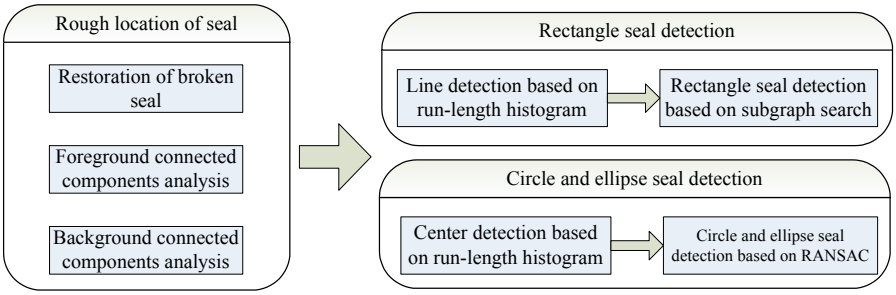


Fig. 1. The algorithm flow chart

2 Rough Location of Seals

If the seal graphic in the document images is imperfect, the noise degradation and the image restoration can be processed first to ensure the high-quality extraction of the target seal before locates the seal graphic.

Since the seal is larger than the character in document images, small connected components such as marks can be eliminated by the connected component analysis in the foreground. For the connected blank region, if the size of the region is too large, the region obviously does not belong to the internal structure of the seal graphic. Then the connected component including this region should be divided. So the connected component can be filtered again in the image background.

Obviously, the above method can be used to get the candidate connected components which possibly include the seal graphic. While the seal graphic has multiple-parameters, such as the rectangular seal and the elliptical seal, this filter method is special useful since the seal can provide more constraints.

3 Accurate Seal Localization Based on the Shape Analysis

3.1 Run-length histogram computation

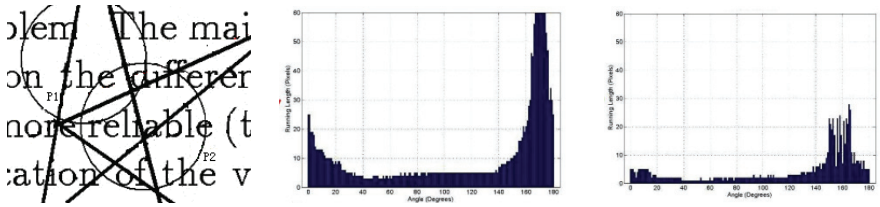
The run-length histogram is an important feature of the binary image. It represents the run-length of the pixel in different directions, and includes the information about width and direction of the straight line which crosses the pixel. Therefore, the run-length histogram is really helpful to identify the straight line.

The run-length of the pixel is defined as the number of consecutive black pixels along one direction. To show the local information of the pixel, the run-length histogram can be acquired by the calculation of the run-length of the pixel along several directions. Figure 2(a) shows a binary image. The run-length histogram of the pixel P1 and the pixel P2 can be calculated by (1) and (2). Figure 2(b) is the run-length histogram of the pixel P1. It shows that the run-length is quite long in line direction if the pixel belongs to the line. Figure 2(c) is the run-length histogram of the pixel P2. It shows that the longest run-length is tangent direction of the curve if the pixel belongs to the curve. Generally, the width and the direction of the straight line which crosses the pixel can be obtained from the shortest run-length and the longest run-length in the histogram. For the run-length histogram, define the run-length feature as followed.

$$WL = \min_{1 \le i \le n} H(i) \tag{1}$$

$$DL = \arg\max_{1 \le i \le n} H(i)n \tag{2}$$

WL includes the width of the straight line and DL includes the direction of the straight line. $H(i)$ is the run-length histogram and n is the number of run-length directions.



(a) (b) (c)

Fig. 2. Computation of run-length histogram of point P1 and P2. (a) a binary document image, and (b) and (c) are the run-length histogram of point P1 and P2.

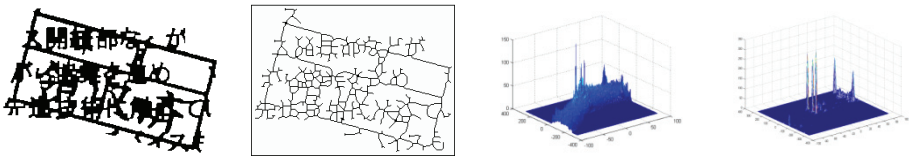
3.2 The detection of the rectangular seal

3.2.1 Improved Hough transform method for line segment detection

Since the seal graphic often overlaps characters and tables, the skeletonized binary image after the still has noises and these noises will definitely disturb the line detection. Note that the point of the skeleton in line has quite long run-length in line direction and the point of the skeleton in burr does not have this feature. Consequently, the feature of the run-length can be used to help the Hough detection of the straight line.

The feature of run-length of the point in the skeleton is calculated first. Then, in the Hough transform, the points which have shorter run-length than the threshold do not vote and the other points only vote in a small angle range along the direction of the long run-length. The above threshold can be determined by the size of the seal. Therefore, the effect of the noise to the Hough voting map can be reduced greatly and the accurate detection of the straight line can be acquired. The voting map of the improved Hough transform can be calculated by (3). (x, y) represents the potential point on the straight line, the $\theta_{\max}$ is the direction of the longest run-length of the point and α is the small angle range of voting, the A is voting accumulator. Figure 3(d) shows the voting map of the Hough transform based on the feature of the longest run-length. Obviously, the improved Hough transform can greatly reduce the invalid voting and sharp the peak of the voting map.

$$\begin{cases} A(\theta,\lambda) = A(\theta,\lambda) + 1 \\ \lambda = x\cos\theta + y\sin\theta, \theta \in [\theta_{\max} - \alpha, \theta_{\max} + \alpha] \end{cases} \quad (3)$$

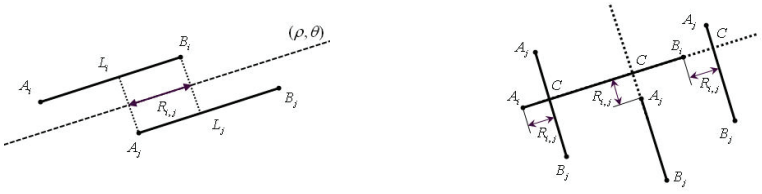


(a) (b) (c) (d)

Fig. 3. Fig. 3 Line detection of Hough transform base on run-length histogram. (a) is a seal image overlapped with words, (b) is the seal's skeleton, (c) is the standard Hough voting accumulator on the skeleton, (d) is our method's voting accumulator on the skeleton.

3.2.2 Neighborhood matrix of the line segments

The neighborhood matrix of the straight line can be used to determine the straight line in the binary image. Suppose $\{R_{ij}\}_{1\le ij\le n}$ is the neighborhood matrix of n straight lines in the binary image. The element R_{ij} in the matrix represents the space position of line L_i and line L_j, such as vertical and parallel.



(a) (b)

Fig. 4. The parallel and vertical relationship of lines. (a) shows the relationship computation of two parallel lines and (b) shows the relationship computation of two vertical lines

Assume the straight line L_i can be defined as a group of four parameters $(\rho_j, \theta_j; A_j, B_j)$. If $\theta_i = \theta_j = \theta$ and $i > j$, R_{ij} is the length of the overlapping projection of the two parallel lines, as shown in Fig.4(a). The value of R_{ij} shows whether or not the two straight lines can be the opposite sides of the rectangle.

If $\theta_i = \theta_j + \pi/2$ and $i < j$, R_{ij} is the distance between the endpoint of the straight lines and the cross point of the two vertical lines, as shown in Fig.4(b). If the two line really intersect, the value of R_{ij} is positive, or it is negative.

The neighborhood matrix of the straight line can be acquired by calculating the space position of each pair of the straight lines. Now the problem of the detection of the rectangular seal can be changed into the problem of the search of the subgraph. We choose the initial vertex first, alternatively search the straight line which is vertical or parallel with the line and decide whether or not these straight lines can form the rectangle which satisfies the constraint condition. If they really satisfy the constraint condition, then storage the rectangle and search the next vertex. If they do not satisfy the constraint condition, search the next vertex directly.

3.3 Detection of the circular and elliptical seal

Obviously, the structure of circular and elliptical seal is centrosymmetric and each pair of symmetry points will vote for the central position. However, when there are too many points in the image, how to select the pair of points efficiently is an important problem. If all the pair of points is selected, the computational complexity is too large. Moreover, the regional peak of the voting map will be hard to be acquired.

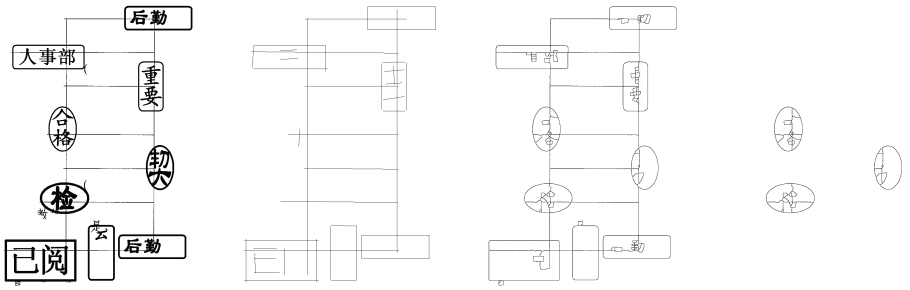
The run-length histogram can be use to select the pair of points in the binary document image.

(1) Select the connected component; compute the run-length histogram of the selected connected component first and then the information of width and direction of each point can be acquired. At last, extract the skeleton of the connected component.

(2) Eliminate the independent line in the background; when there are some lines in the background of the binary image, the voting for the central position of the circle and the ellipse will be greatly weaken by the voting of the point in lines. These lines

can be eliminated by select the connected component in the skeleton. We detect the line in the connected component and then study the two adjacent white backgrounds. If the common boundary of them is a long straight line, the boundary should be eliminated and the two backgrounds should be combined. These operations will be repeated for several times unless all the lines have been eliminated. After the above step, the skeleton will be selected again based on the background and the foreground and eliminate the connected region which is too large or too small. Fig. 5 shows the elimination of the lines in the skeleton.

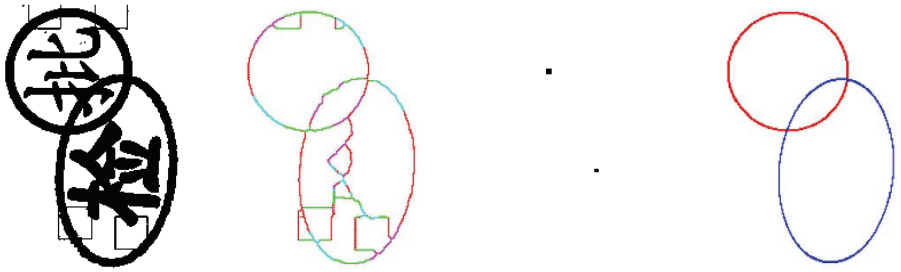
(3) Voting for the central position; for two points which have the same direction and width, if the distance between the two points satisfy the constraint of the seal, then calculate the central position based on the two points. Voting in the corresponding position in the array and the central point of the circle or the ellipse will be finally determined based on the array.



(a) (b) (c) (d)

Fig. 5. Delete the long line of the candidate connected component, (b) is the line detection result of (a), (c) is the skeleton of (a) and (d) is the result after deleting line

(4) Calculation of the central point; the central point of the circle and the ellipse can be determined by the regional extreme value of the array. However, this method may causes some errors when there are several seal graphics or much noise in the connected component. So an improved method is proposed to determine the central point. First, we calculate the run-length histogram of all the points in the skeleton along the four directions {0°,45°,90°,135°}. Then the points in the skeleton can be divided into four subsets. For each subset, the array can be acquired by the voting for the central point. Since both circle and ellipse are centrosymmetric, there will be a regional extreme value in the central position of the circle or the ellipse. Check whether or not there is a regional extreme value in the same position of the four arrays. The real central position can be determined based on the above method. Compared with the original method, the improved method does not need the threshold and it can also be used in the connected components which have several seals. Fig. 6 shows the example that how to determine the central position of the connected component which includes the circle and the ellipse.



(a) (b) (c) (d)

Fig. 6. The center location of circle and ellipse. (a) is a candidate connect component, and (b) is the stroke direction map, (c) shows the center's position，(d) is the circle and ellipse detection result of RANSAC fitting

After the determination of the central position of the circle or the ellipse, we still need to calculate some other parameters. For the circle, the radius should be calculated, and for the ellipse, the major and the minor axis should be calculated. The RANSAC method is used to calculate these above parameters because this method is robustness when the point set has many exterior points. The pair of points which voting for the central position is determined first and this operation can be done by scan the skeleton again. Figure 6(d) shows the result of the RANSAC method for the circle and the ellipse in figure 5.

4 Experiment Results

The experiment is conducted in MATLAB 6.0 and the average computation time for each document image is obtained from the experiments of 200 200dpi document images and 100 400dpi document images on a 2.8GHz Intel Pentium IV PC with 1GB memory. There are several seal graphics with different shapes in the experimental document images. Each experimental document image for the circular seal has four to eight circular seals. Each experimental document image for the elliptical seal has two to five elliptical seals and each experimental document image for the rectangular seal has eight to twelve rectangular seals. These seals may overlap other objects in document images, such as characters, tables and graphics.

The seal graphic with different shapes can be reliably detected with the high recall and precision by the proposed method in this paper. The average computation time is 20s for a 200dpi document image. The computation time may be reduced while the algortihm is optimized after the futher research.

Table 1. The performanc statistic of precision, recall and processing speed

Resolution & Size	Circle seal			Ellipse seal			Rectangle seal		
	P	**R**	**T**	**P**	**R**	**T**	**P**	**R**	**T**
200dpi,A4	98.9%	100%	18.8s	92.4%	100%	20.3s	98.9%	95.9%	16.5s
400dpi,A4	98.9%	100%	20.0s	91.6%	100%	22.7s	98.9%	96.4%	18.1s

Note: P-precision, R-recall, T-the average processing time. T is related with the number of seals in the test image.

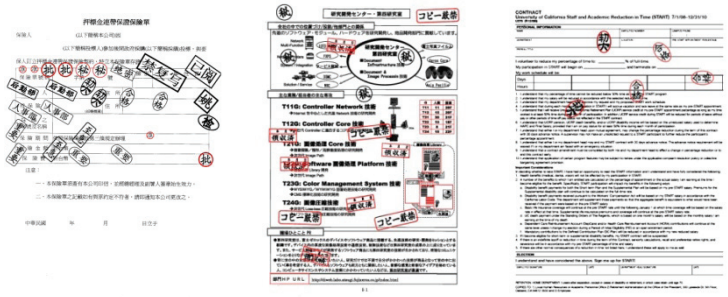


Fig. 7. The result of kinds of seals detection.

5 Conclusion

This paper discussed a novel method for seal detection and extraction. This method analyzes connected components in both foreground and background, filter connected components based on the constraint of the range of the seal graphic. According to the run-length histogram feature, the lines and central position of seal graphic can be located fine. Finally, plenty of tests were conducted to show the validity of the method. The results with high recall and precision showed the fact that this method can detect and identify the seal graphic with specified shape in document images and the method is definitely a practical algorithm to detect the seal graphic in the mobile devices in the future.

Acknowledgment. This work was supported by the NSF of China (No.60775017, and 90920008).

References

1. McLaughlin R.A.:Randomized Hough transform: improved ellipse detection with comparison. In: Pattern Recognition Letters, vol. 19, no. 3-4, pp.299-305 (1998)

2. Xu L., Oja E., Kultanen P.: A new curve detection method: randomized Hough transform (RHT). In: Pattern Recognition Letters, vol. 11, no. 5, pp. 331-338 (1990)
3. Ziqiang Li: Generalized Hough Transform: Fast Detection for Hybrid Multi-Circle and Multi-Rectangle. In: the 6th World Congress on Intelligent Control and Automation, pp. 10130-10134 (2006)
4. Cheng Y. C., Lee S.: A new method for quadratic curve detection using K-RANSAC with acceleration techniques. In: Pattern Recognition, vol. 28, no. 5, pp. 663-682 (1995)
5. Fitzgibbon A., Pilu M., Fisher R. B.: Direct least square 2tting of ellipses. In: IEEE Trans. on PAMI, vol. 2l, no. 5, pp. 477-480 (1999)
6. Yin P.: A new circle/ellipse detector using genetic algorithms. In: IPPR on CVGIP, pp. 362-368 (1998)
7. Ueda K., Nakamura Y.: Automatic verification of seal impression patterns. In: 7th. International Conference on Pattern Recognition, pp. 1019-1021 (1984)
8. Fan T., Tsai W.: Automatic Chinese seal identification. In: Comput. Vision Graphics Image Process., vol. 25, pp. 311-330 (1984)
9. Chen, Y. S.: Automatic identification for a Chinese seal image. Pattern Recognition, 29(11), pp.1807-1820,1996.
10. Wu Yi, Chiang J., Wang Ruiching: Seal Identification Using the Delaunay Tessellation. In: Proc. Natl. Sci. Counc. ROC(A),Vol. 22, No. 6, pp. 751-757 (1998)
11. Zhu G, Jaeger S, Doermann D.: A robust stamp detection framework on degraded documents. In: Proceedings of SPIE Conference on Document Recognition and Retrieval XIII, pp.1-9 (2006)
12. Yangxing Liu, Takeshi IKENAGA, et al.: A Novel Approach of Rectangular Shape Object Detection in Color Images Based on An MRF Model. In: 5th IEEE Int. Conf. on Cognitive Informatics (ICCI'06), pp.386-393 (2006)

Tracking Fingers in 3D Space for Mobile Interaction

Shahrouz Yousefi, Farid A.Kondori, and Haibo Li

Digital Media Lab., Applied Physics and Electronics, Umeå University
SE-901 87, Umeå, Sweden
{shahrouz.yousefi,farid.Kondori,haibo.li}
@tfe.umu.se
http://www.tfe.umu.se/english/research/dml/

Abstract. Number of mobile devices such as mobile phones or PDAs has been dramatically increased over the recent years. New mobile devices are equipped with integrated cameras and large displays which make the interaction with device easier and more efficient. Although most of the previous works on interaction between humans and mobile devices are based on 2D touch-screen displays, camera-based interaction opens a new way to manipulate in 3D space behind the device in the camera's field of view. In this paper, our gestural interaction relies heavily on particular patterns from local orientation of the image called *Rotational Symmetries*. This approach is based on finding the most suitable pattern from the large set of rotational symmetries of different orders which ensures a reliable detector for fingertips and human gesture. Consequently, gesture detection and tracking can be used as an efficient tool for 3D manipulation in various applications in computer vision and augmented reality.

Key words: Mobile interaction, Rotational Symmetries, Gesture detection, Tracking

1 Introduction

Nowadays gesture detection, recognition or tracking are terms which have frequently been encountered in discussions of human computer interaction. Gesture recognition enables humans to interact with computers and makes the input devices such as keyboard, joystick or touch screen panels redundant.
New mobile phones with higher camera resolution provide the opportunity to take advantage of image analysis techniques to utilize in different applications. Our experiments on mobile applications reveal that in 2D interaction on the screen, users have limitations in moving in depth, zooming in to observe the details of an image, map, text or zooming out to skim through a data. Even in more complicated applications rotations around different axes are unavoidable. Here the question is whether it is possible to solve these limitations by introducing a new interaction environment? Taking advantage of three-dimensional space

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

behind the mobile device can remedy these problems. The retrieved gestural information under the mobile phone's camera in 3D space help us to implement many applications. In this paper we discuss how gesture detection and tracking under the mobile phone's camera can be used in a novel way for human mobile phone interaction. Finally, we introduce applications which clarify the use of our discussion in reality. Our gesture recognition method is based on the *rotational symmetries* detection in video input from the camera. This method finds patterns from local orientation of the image. The implemented operator searches for particular features in local images and detects expected patterns associated with human gesture. Tracking the detected gesture enables humans to interact with mobile phone in 3D space and manipulate in various applications. This method is based on very low-level operations with the minimum level of intelligence which makes the system more reliable and efficient in gesture detection and tracking part.

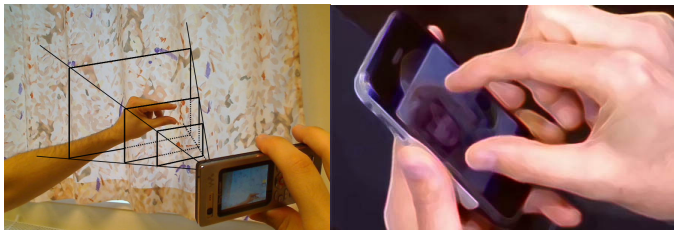
2 Related Work

Designing a robust gesture detection system using a single camera independent of lighting conditions or camera quality is still a challenging issue in the field of computer vision. A common method for gesture detection is marker-based approach. Most of the augmented reality applications are based on marked gloves for accurate and reliable fingertip tracking [4, 5]. However, in marker-based methods users have to wear special inconvenient markers. Moreover, some strategies rely on object segmentation by means of temperature [6]. Robust finger detection and tracking could be gained by using a simple threshold on the infrared images. Despite the robustness, thermal-based approaches require expensive infrared cameras which are not provided to mobile devices. Other approaches such as template matching [7] and contour-based methods [8] have been employed for special hand gesture detection, though they are generally computationally expensive. Another set of methods for hand tracking are based on color segmentation in appropriate color space [9]. What we present in this paper is simply based on low level operators, detecting natural features without any effort to have an intelligent system. We propose a way to take advantage of lines, curvatures, circular patterns and in general rotational symmetries associated with the model of the human fingers which leads to the detection of the human gesture.

3 Mobile Interaction in 3D Space

Having more effective interaction with mobile phone could be the most significant reason behind the manufacturing of mobile phones with larger screens during the recent years. Although the idea of working with larger touch screen displays helps users to have a better interaction with device, their limitations in 2D space manipulation remain an unsolved issue. In addition, mobile devices with larger screens are heavier and not easy to carry and normally people would like to buy a pocket-size mobile phone. A novel solution for limitations of 2D touch-screen

displays is taking advantage of 3D space behind the mobile phone's camera. Manipulation in the camera's field of view provides a chance for user to work with any mobile phone regardless of the screen size or touch sensitivity. As it is shown in Figure 1 the user's hand in farther distances from the camera occupies smaller place in the screen which is a positive point to compensate the limited area for fingers on 2D displays. Moreover, behind the camera users are capable of moving in depth which could handle a lot of difficulties in various applications.



(a) 3D interaction (b) 2D interaction

Fig. 1: (a)User capability to move in depth and change the finger size according to the distance to the camera. (b)Limited area for user's fingers in 2D interaction.

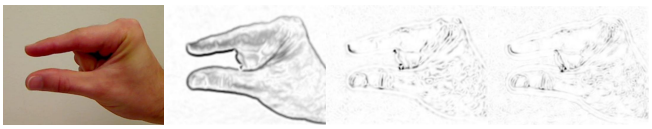
Fig. 2: System Overview

4 Rotational Symmetries and Pattern Recognition

Rotational Symmetries are specific curvature patterns detected from local orientation [3]. This theory was developed in 80's by Granlund and Knutsson [1]. The main idea behind that is to use local orientation [3] to detect complex curvatures in double-angle representation [3]. Using a set of filters on the orientation image will result in detection of number of features in different orders, such as curvatures, circular and star patterns.

The idea of taking advantage of the rotational symmetries in gesture detection is rather general and complex at first look, but modeling of the fingers by the choice of the rotational symmetry patterns of different classes could lead us to differentiate between fingertips and other features even in complicated backgrounds. Since the natural and frequently used gesture to manipulate objects

in 3D space is similar to Figure 3.a, this model can satisfy our expectations for different applications.



(a) Test image (b) 0:th order (c) 1:th order (d) 2:th order

Fig. 3: Response of the test image to different orders of rotational symmetries

As previously discussed, each set of rotational symmetries represents particular patterns. Our experiments based on different test images of different scales and backgrounds reveal that fingertips have significant response to the group of first order rotational symmetries. This group represents curvatures in different phases (see Figure 3). Our observation shows that searching for the first order rotational symmetries in image frames with suitable filter size will result in high probability of responses of fingertips in different scales. For example, searching the image of the human gesture by the first order detector and increased selectivity grants our expectations up to this level.
In order to increase the selectivity of the rotational symmetries of our interest, different methods have been developed. What we introduce here is normalized inhibition [3]. Consider that we aim to detect the rotational symmetries of the order n (1 :th order in our case), but not the order k(0 :th order). Then the normalized inhibition is computed as,

$$\check{s}_n = \langle a_n b_n, z\rangle \prod_k \left(1 - \frac{|\langle a_k b_k, z\rangle|}{\langle a_k, |z|\rangle}\right) = s_n \prod_k \left(1 - \frac{|s_k|}{\langle a_k, |z|\rangle}\right). \quad (1)$$

Where z is the double angle representation of the image and

$$a_n(\mathbf{x})e^{in\varphi} = r^{|n|}g(\mathbf{x})e^{in\varphi} = \begin{cases} (x+iy)^n g(x,y) & n \geq 0 \\ (x-iy)^{-n} g(x,y) & n < 0 \end{cases} \quad (2)$$

5 Gesture Detection And Finger Tracking

Video frames captured by the camera are analyzed to detect the particular gesture. Then the fingers are tracked in the coming video frames and their updated positions will be utilized in various applications.

5.1 Gesture Detection

The gesture detection algorithm relies on the first order rotational symmetries responses. As it was mentioned before, fingertips respond to the first order rotational symmetries substantially. On the other hand, complex background regions

could have high probability of response to this group of rotational symmetries, which can cause incorrect detections (see Figure 4).

(a) (b)

Fig. 4: Detection affected by complex background. (a)Test image,(b)Black stars, incorrect detections caused by background, white stars, the correct detections.

Thus, we ensure that fingertips are correctly detected by means of three criteria; phase and magnitude of the response to the first order rotational symmetries and the human skin color (see Figure 5).

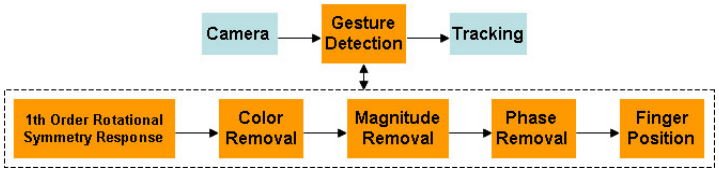


Fig. 5: Gesture detection flow

The first step to detect the fingertips is to remove the features with different values from the human skin color. RGB values of the detected pixels are compared to the human skin color and the candidate pixels will be considered for further steps. The next step is to compute the magnitude of the rotational symmetries responses. Recall from our previous discussions, the gained magnitude should be high for every candidate pixel to be considered as a fingertip. A constant value is set (according to [3]) to the magnitude of the response. Eventually, the last criterion that should be under consideration is the phase of the rotational symmetries responses.
Based on various experiments, we discovered that fingertips have great response to the first order rotational symmetries with phase=π. Hence, we have a critical clue to remove unexpected responses.

5.2 Localization by Clustering

We might gain several pixels around the fingertips after gesture detection. Therefore, we construct clusters [2] from classified data and estimate the center point

of the clusters as the best approximation of the fingertips positions in the image. Figure 6 illustrates the result after three steps of removing unexpected responses.

(a) (b) (c) (d)

Fig. 6: Three steps of removing unexpected responses. (a) original image, (b)after color skin removal, (c) after magnitude removal, (d) after phase removal

5.3 Finger Tracking

After fingertip detection, the main challenge is to track them in the consecutive frames. Thus, we define a searching box around them for localization. In this way, we know that in the next coming frames the fingertips are expected to be in the pre-defined searching box and instead of searching the whole image, which is computationally expensive and time consuming, small patch of the image will be processed.

6 Experimental results

Our observations on different experiments revealed that in order to have a robust detection and consequently tracking, our system should be to some extent scale and rotation-invariant to the human's gesture. As a matter of fact, for a particular gesture behind the mobile phone's camera, users have freedom to move in a reasonable distance. Moreover, depending on the application, they are free to rotate in different angles. Our experiment indicates that the effective interaction happens in the area between 15-25cm from the camera. Interaction in the area beyond 25cm does not seem to be convenient for users. Clearly, for distances below 15cm, gesture occupies a large area in the screen and degrades the interaction. The first version of our algorithm had been implemented and tested in Matlab 7.5.0, but due to the fact that we aim to benefit from this approach in real-time applications, finally, we implemented the software in C++. Figure 7 illustrates the system performance in the tracking of the particular curve on a complex background. In this example the user is asked to follow the defined curve drawn on the screen. Red circles mark the position of the detected gesture corresponding to each image frame. The error in the tracking of the original curve for more than 200 frames is plotted in Figure 8. The mean value of the error (6.59 pixels) shows the slight difference between the original curve and the one plotted by the tracked gesture which is quite satisfying.

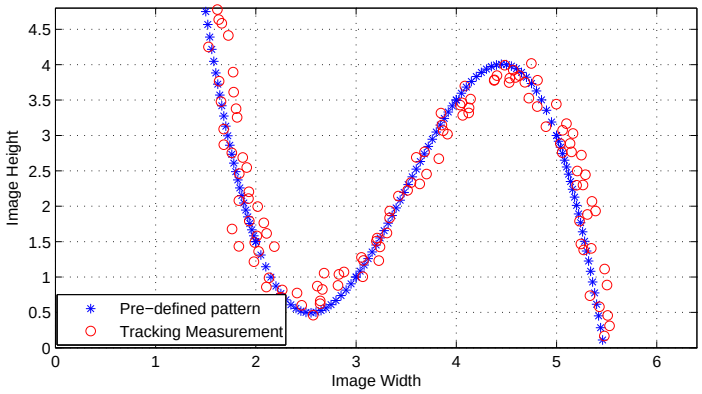


Fig. 7: Example of the system performance in gesture tracking

7 Applications

Here we introduce some applications which could gain from our interaction approach.

7.1 Map and Text Reading

Current applications of text reading or map observation in mobile phones are not convenient for users, particularly if they observe in z-axis to zoom in or zoom out. As we might experience, in new mobile devices applications use double-finger operation on touch panels to move in depth but gesture-based motion in any direction and also in depth is the ideal case in compare with the current technologies.

7.2 Photo Edit Toolbox

Many interesting applications in mobile devices are related to multimedia. Photo or video edit applications can benefit from gestural interaction. Photo extraction, photo attachment, visual effects or other tools will be more user-friendly by 3D interaction.

7.3 Augmented Reality, 3D Object Manipulation on 2D Screen

The most exciting applications of gesture detection could be implemented in the field of augmented reality. For instance, in 3D object manipulation, users can segment objects by their gesture. Afterwards, the model of the segmented object can be extracted from the database with respect to shape, size and color. In the same manner of gestural interaction, this model can be resized, reshaped and rotated in any direction for any purpose.

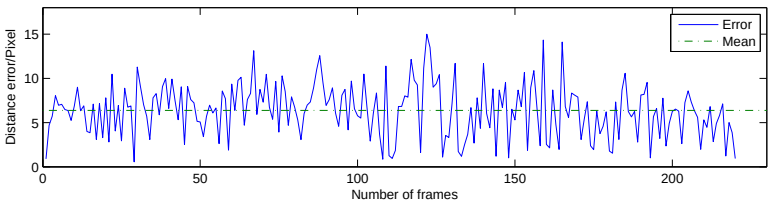


Fig. 8: Error of the tracking in a sequence of images

8 Conclusion

In this paper we presented a novel approach for 3D camera-based gesture interaction with mobile devices. The detection algorithm can estimate the position of the fingertips and tracking part updates the positions in consecutive frames. Robustness and simplicity are the main advantages of this approach that rely on the low level operations and equipment-free gestural interaction. Our detection and tracking algorithms are computationally efficient and work well in practice. Although wrong detections caused by quite complex backgrounds are rare, but future work can concentrate on improvement and optimization of the algorithm.

References

1. Knutsson, H., Granlund, G.H.:Apparatus for determining the degree of variation of a feature in a region of an image that is divided into discrete picture elements., US-Patent 4.747.151, 1988. Swedish patent, 1986.
2. Moore., A.:K-means and Hierarchical Clustering-Tutorial Slides, `http://www-2.cs.cmu.edu/~awm/tutorials/kmeans.html`
3. Johansson., B.:Low Level Operations and Learning in Computer Vision, *Linkoping Studies in Science and Technology Dissertation No. 912*, Department of Electrical Engineering,Linkopings universitet, SE-581 83 Linkoping, Sweden Linkoping 2004.
4. Dorfmueller-Ulhaas, K., Schmalstieg., D.:Finger Tracking for Interaction in Augmented Environments. In: 2nd ACM/IEEE Int'l Symposium on Augmented Reality, pp. 55-64, 2001.
5. Mggioni, C.:A novel gestural input device for virtual reality. In: Virtual Reality Annual International Symposium, IEEE, 1993, pp. 118 - 124.
6. Iwai, D., Sato, K.:Heat Sensation in Image Creation with Thermal Vision. In: ACM SIGCHI Int'l Conf. on Advances in Computer Entertainment Technology, pp. 213-216, 2005.
7. Rehg, J., Kanade., T.:Digiteyes Vision-based Human Hand Tracking. Technical Report CMU-CS-TR-93-220, Carnegie Mellon University, 1993.
8. Zhou., H., Ruan., Q.:Finger Countour Tracking Based on Model. In: Conf. on Computers, Comunications, Control and Power Engineering, pp. 503-506, 2002.
9. Bencheikh-el-hocine, M., Bouzenada, M., Batouche, M.C.:A New Method of Finger Tracking Applied to the Magic Board. In: Int'l Conf. on Industrial Technology, pp. 1046-1051, 2004.

People, places and playlists: modeling soundscapes in a mobile context

Nima Zandi, Rasmus Handler, Jakob Eg Larsen, and Michael Kai Petersen

Technical University of Denmark,
DTU Informatics, Cognitive Systems Section
Richard Petersens Plads, Building 321,
DK-2800 Kgs. Lyngby, Denmark,
{jel|mkp}@imm.dtu.dk

Abstract. In this paper we present an initial study of music listening patterns on mobile devices combined with contextual information. The study included N=7 participants that carried a smart phone for a duration of two weeks. The participants used the main features of the phone along the music player capabilities. All phone activities and data from embedded sensors were recorded along the music being played on the device. We report initial indications that listening patterns in terms of music genre preferences are influenced by whether the user is in a static environment or on the move. Applying a simple decision tree algorithm to identify what contexts determine the preferences indicate that our listening patterns change over time, suggesting that music applications utilizing context information must be designed to adapt to our shifting preferences as they continuously evolve.

Keywords: Mobile, context-awareness, music, genre, listening pattern

1 Introduction

The aspects of context-awareness and context-aware applications have gotten much attention for more than a decade. In this paper our focus is on the consumption of music content in mobile scenarios involving multiple contexts. It is our assumption that the way people access and use content is highly dependant on the particular context. The context include the time, place, tasks, motivation, and the history of the interaction.

When people carry their mobile phones throughout the day it provides a unique opportunity to capture contextual information from the wide range of embedded sensors. Together these sensors provide an interesting source of information about activities, people, places and other entities. The advancements of mobile phones has increased the potential for novel context-aware mobile applications. Present off-the-shelf smart phones have several embedded sensors, such as, GPS, accelerometer, light sensor, proximity sensor, microphone, camera, as well as multiple network connectivity options, such as, GSM, WLAN, and Bluetooth. As such, the mobile phone can potentially serve as a proxy in terms of providing information about the context of the human user [1].

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

In a recent study Song et al. [2] discuss the predictability of human behavioral patterns based on detected movement patterns using GSM cellular information for location approximation. The interesting result is a predictability as high as 93%, which provides an indication of the potential for future context aware mobile applications. Mobile applications involving music has also gained interest recently. Especially novel ways of navigating music collections is available in the commercial space, examples include Moodagent for the iPhone and Playlist DJ for Symbian, which allows the user to navigate a music collection in terms of mood, rhythm and style of the music. We hypothesize that contextual information obtained from a mobile device can offer useful information in terms of providing additional input for music recommendation for the individual and in a social context.

2 Context

In order to capture contextual information on mobile phones in our study, we utilize an existing solution – Mobile Context Toolbox – available for the Symbian S60 mobile operating system [1]. The system is built in multiple layers (as depicted in Fig. 1) on top of the Nokia S60 platform using Python for S60 (PyS60) with a set of extensions for accessing low-level sensors and application data.

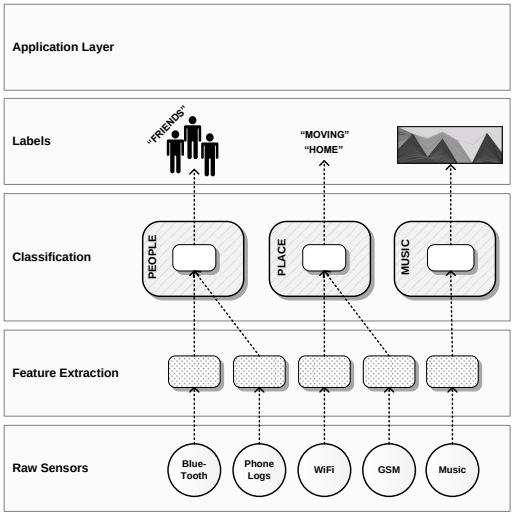


Fig. 1. Mobile Context Toolbox architecture

The underlying toolbox provides interfaces for accessing multiple low-level sensors, encapsulated in higher-level adapters. The inspiration for the layered approach is the framework described by Salber et al. [3] where the emphasis is on a clean cut between system resources, inferring contextual information and applications using it. Thus the details of the invidual sensors are abstracted

to infer higher-level contextual information and focus on the feature extraction and classification of the obtained sensor data. Taking this approach simplifies the process of aggregating sensor data into higher-level contexts, as-well as making the framework extensible and adaptable to changes in the underlying platform.

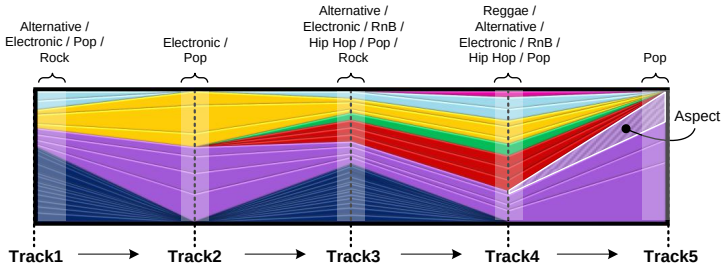
In the present study we have focused on obtaining contextual information about people, places and music. As shown in Fig. 1, information is obtained from the Bluetooth sensor and phone log in order to extract features related to people. Wi-Fi and GSM cellular information is used to extract features about the present place. Finally, the existing Mobile Context Toolbox has been extended with a virtual sensor obtaining information from the embedded music player application on the phone. Thus it acquires information about tracks being played, that is, extract music features including song title and artist information obtained from the embedded track metadata. Each of these features are translated into meaningful labels, and finally the application layer can utilize the contextual information inferred from the system by means of these contextual labels.

3 Music

The track information from the music player is incorporated as yet another sensor. The aim is to model not only the people and places as our mobile context, but also the constantly changing frame of mind reflected in the music we listen to. Although the artist and genre descriptions associated with the tracks as ID3 tags might be limited to terms such as 'pop', we are able to extend this based on available data from music social networks like *last.fm*. Tag-clouds generated by hundred thousands of users, might be interpreted as high dimensional representations of artist information, musical genres as well as the perceived emotional context of songs. Studies have shown that people often tend to agree not only on what emotional tags to use but also what tracks they are applied to [4]. The genres which form tag-clouds are far from crisp categories, but have large overlaps between what might vaguely be termed 'rock' versus 'pop', or could be labeled 'indie' in contrast to 'alternative'. When applying probabilistic latent semantic analysis to extract the underlying topics behind the most frequently co-occurring words, aspects like 'pop' will appear multiple times in different contexts coupled with tags such as 'soft, rock' or 'love, romantic' while emotions such as 'chillout' would be fused together with other tags like 'electronica, ambient, downtempo' [5]. This allows us to build track and session signatures based on the ID3 tags that are enhanced to capture the underlying semantic aspects of the music. Initially defining track signatures for the songs that are being played, each $Track_i$, $i = 1, \cdots, m$, consists of aspects $\{A_1, ..., A_{91}\}$ where A_j is based on tags $\{T_{j1}, ..., T_{jn}\}$. For every j'th aspect, A_j, in $Track_i$, the ratio of tags is calculated and combined to obtain a track signature for the i'th track. Subsequently we collect tracks into sessions if the user has been listening to at least 3 songs in a row, by combining the above track information into a session signature

$$Session_l^{Signature} = \{\frac{1}{M_l}\sum_{i=1}^{M_l} AspectHitRatio_{1i}, ..., \frac{1}{M_l}\sum_{i=1}^{M_l} AspectHitRatio_{Ni}\}$$

constituting an average of the ratios of tag co-occurrences defined in the track signature, where $M_l = \#\{\text{tracks in session nr. l}\}$. Generalizing the tag co-occurences into broader categories of musical style, allow us to define the changing genre characteristics over time, within the sequences of tracks that constitute a playlist (Fig.2). For visual clarity the different colors are assigned to the different genres used. Based on Pearson correlation we are able to define the similarities between different sessions, across multiple users within varying mobile contexts.



(a) Track sequence signature visualization

(b) Genre colors

Fig. 2. An example generalized track sequence signature (a) derived by enhancing the track ID3 metadata with social network tags, that capture the changing genre characteristics over time. Each music genre has a unique color (b) as shown in (a).

.

4 Experiment

Data was acquired from N=7 participants over a duration of two weeks. The participants were asked to use a Nokia N95 8GB smart phone with our Mobile Context Toolbox preinstalled along a collection of MP3 music files. Participants were asked to use the phone as their standard phone (with personal SIM card), and as their MP3 player for the duration of the experiment. Table 1 provides an overview of the contextual data obtained related to places and people for all 7 participants, where * marks the number of unique data points. In addition the total user interaction with the embedded music player application is included. 'Played' refers to the number of tracks that has started playing on the music player, and 'listened' includes the number of tracks where more than half of the track has been played, similar to the criteria used in *last.fm*.

P	BT	BT*	WiFi	WiFi*	GSM	GSM*	Ph	Ph*	Played	Listened	Unique
1	12620	1387	147333	2892	2043	242	343	35	337	160	85
2	10422	475	69973	4013	2792	344	305	35	474	153	100
3	7406	203	63495	777	402	105	92	20	375	190	48
4	3095	546	25375	1087	3516	296	286	28	524	292	68
5	5032	798	89917	2329	1573	175	164	35	173	110	58
6	15127	2675	75948	3915	1985	472	398	41	742	167	124
7	7654	1849	88222	2306	1132	169	113	17	198	94	65
Total	61356		560263		13443		1701		2823	1166	

Table 1. Overview of obtained contextual data for all 7 participants. BT is Bluetooth data, Ph is phone call/sms log, and Played, Listened and Unique refer to music tracks.

Contextual categorization of music playlists were generated over 2 weeks, where each sequence signature has been linked to labels such as 'at home', 'in transition' or 'weekend', inferred from low level location and motion data collected, as shown in Table 2. In essence capturing the contextual listening habits of a user combined with the genre preferences derived from the sequences of tracks that have been played. The upper and lower rows in the table indicate sessions from the first and second week respectively based on data generated by participant P_1.

Nr.	Transition	Home	O.K.P	U.Place	T.O.D	WeekDay	S.Category
1	True	False	False	False	Day	Work	C
2	True	True	False	False	Day	Work	A
3	True	False	False	False	Evening	Work	B
4	True	False	False	False	Day	Work	A
5	True	False	False	False	Day	Work	C
6	False	True	False	False	Evening	Work	A
7	True	False	False	False	Day	Work	D
8	True	False	False	False	Day	Work	E
9	False	False	True	False	Day	Work	A
10	True	True	False	False	Day	Work	A
11	False	True	False	False	Day	Weekend	B
12	True	False	False	False	Evening	Weekend	D
13	True	False	False	False	Day	Work	B
14	False	False	True	False	Day	Work	F
15	False	False	True	False	Day	Work	G
16	False	False	True	False	Day	Work	F
17	True	False	False	False	Day	Work	H
18	True	False	False	False	Day	Work	D
19	False	False	True	False	Day	Work	G
20	False	False	True	False	Day	Work	C

Table 2. Contextual categorization of music playlists generated over 2 weeks for participant P_1. O.K.P is other known place, such as work, U. place is unknown place, T.O.D is time of the day, and S.Category is session category.

5 Results and Discussion

Learning how music genre preferences are associated with labels such as 'at home', 'in transition' or 'other known place', inferred from low level location and motion data continuously collected by the Mobile Context Toolbox, makes it possible to contextually categorize music based on the constantly changing usage scenarios. Even when only considering a limited test based with seven participants forming a small scale network, patterns emerge that indicate how it might be possible to share music by traversing the social graph and find users in a similar context or listening to playlists resembling our own, as illustrated in Fig. 3.

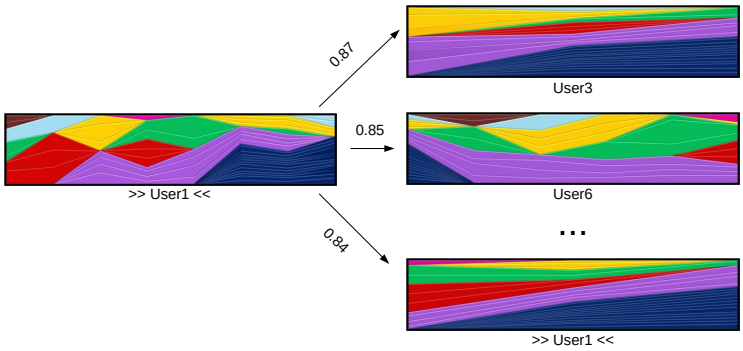
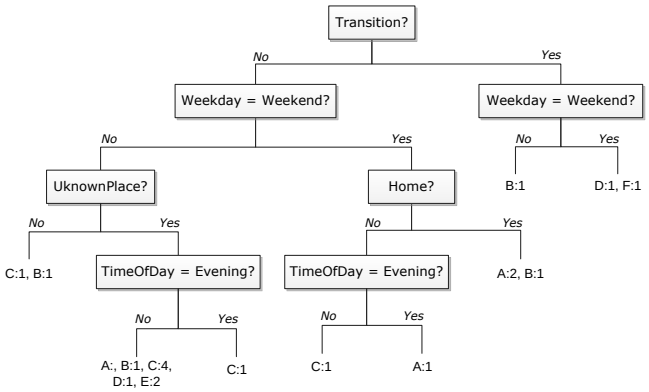


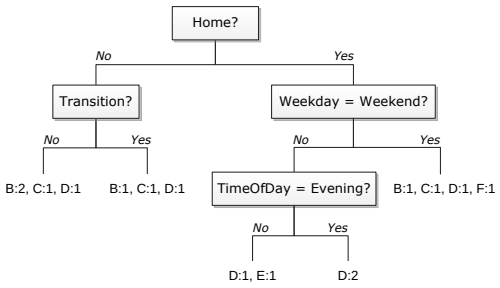
Fig. 3. Traversing the social graph similar music playlists are identified and ranked using Pearson correlation as distance measure based on participant P_1's session signature. The correlated sessions are retrieved from different levels in the social graph, e.g. not only friends have similar sessions but also friends of friends. Here the highest-ranked session is from a friend, whereas the following session is from a friend of a friend.

When applying a simple machine learning algorithm to untangle the user choices that define the contextual genre preferences, it becomes apparent that these structures are extremely dynamic. Not only do the contextual labels that influence what music we listen to differ from one user to another, but also change from one week to another when viewed from the perspective of a single user. This means training a classifier on contextual listening patterns over a short period and use it for future prediction is not likely to work, unless it is continuously adjusted to shifting preferences. Nevertheless clear tendencies in listening patterns appear to emerge for habits associated with contexts corresponding to static scenarios like 'at work' or 'gym' versus what genres are being exploring when on the move. The study by Song et al. [2] use entropy as a measure to assess the degree of predictability in patterns of movement based on extrapolated GSM cellular information for location approximation, which indicates that human behavior may be determinable up to 93%, of the time. However, when minimizing

entropy in our approach, based on a decision tree algorithm that finds the contextual labels which as top nodes best define the genre preferences for each usage scenario, it seems that size matters when selecting a temporal frame. That is, a dynamic approach to prediction appears to be essential once we move beyond determining likely nodes of location and enter the unchartered territory of how our media preferences are influenced by the people and places constituting our constantly shifting mobile context.



(a)



(b)

Fig. 4. Decision trees classifying the user choices that determine the associations between contextual labels and playlist genre characteristics, trained based on data generated in the first (a) and second (b) week respectively for participant P_4. Splitting the data on a weekly basis highlights the dynamic character of the contextual listening patterns. The top nodes corresponding to the conditions that are most significant for defining the correlations between mobile contexts and genre preferences differ not only between users, but also change on an individual basis within the two week period.

Example decision trees for participant P_4 is provided in Fig. 4. In the decision tree the nodes are the spatiotemporal context labels generated by the Mobile Context Toolbox, whereas the leafs at the end of branches define the preferred genre categories corresponding to each usage scenario. The algorithm finds the variables that best divide the data, meaning that the nodes which are pushed towards the top represent the conditions which in terms of information gain contribute the most to explaining the underlying structure.

6 Conclusions

Although our study is based on a very limited number of users constituting a small scale social network, our initial findings suggest that it is possible to define contextual categories linking our music genre preferences to labels continuously inferred from low level location and motion data generated by the Mobile Context Toolbox running on the smart phones. The study has supported our expectations that listening patterns in terms of preferred music genres are influenced by conditions defining whether we are in a static environment or on the move. Applying a simple decision tree algorithm to identify what contextual labels determine music preferences, our results indicate that our listening patterns are continuously transformed over time. This indicates that even though we may observe distinct tendencies in habits related to the underlying context, future recommender systems must allow the application to adapt to our changing music listening patterns as they are influenced by context but also appear to continuously evolve over time.

Acknowledgments The authors would like to thank the 7 participants which took part in the experiments. Also thanks to Nokia Denmark and to Forum Nokia for the equipment used in the experiments.

References

1. Larsen, J., Jensen, K.: Mobile Context Toolbox-an extensible context framework for S60 mobile phones. In: Smart Sensing and Context: 4th European Conference, EuroSSC 2009, Guildford, UK, September 16-18, 2009. Proceedings, Springer-Verlag New York Inc (2009) 193
2. Song, C., Qu, Z., Blumm, N., Barabasi, A.: Limits of predictability in human mobility. Science **327**(5968) (2010) 1018
3. Salber, D., Dey, A.K., Abowd, G.D.: The context toolkit: aiding the development of context-enabled applications. In: CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems, New York, NY, USA, ACM (1999) 434–441
4. Hu, X., X, M.B., Downie, S.: Creating a simplified music mood classification ground-truth set. In: Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR). (2007) 309–310
5. Levy, M., Sandler, M.: Learning latent semantic models for music from social tags. Journal of New Music Reseach **37**(2) (2008) 137–150

Recognition-Based Error Correction with Text Input Constraint for Mobile Phones

Zhipeng Zhang, Yusuke Nakashima and Nobuhiko Naka

Research Laboratories, NTT DOCOMO, INC

3-6 Hikari-no-oka, Yokosuka, Kanagawa, Japan

zzp@nttdocomo.co.jp

Abstract. We propose a highly practical error correction method for mobile phones. Given a recognition error, we use a re-recognition-based correction method that is constrained by the user's input of the first few correct characters. It forces the system to produce text that begins with these characters. Since this method refreshes the correction result in response to each character input by the user, error correction is finished with the fewest possible keystrokes. Our experiment shows that this method reduces input stroke quantity by 70%; its correction rate reaches 80%.

Keywords: Speech recognition, Error correction, Usability

1. Introduction

The demand for speech-based interfaces for mobile applications such as speech translation and e-mail continues to increase. However, mobile devices have limited computation, memory, and energy resources. Thus they have difficulty in performing complex speech recognition tasks. Distributed Speech Recognition (DSR) [1] provides a practical framework for mobile phone speech recognition. Despite all efforts at improving the recognition accuracy, some recognition errors are inevitable, so an error correction method that suits mobile devices is vital in making the mobile phone speech-interface and applications more popular.

Several error correction methods [2-5] have been studied. The multi-modal based error correction method [2] that combines with pen-based input reduces the key operations at the cost of additional devices, which makes it difficult to implement this method on mobile phones. It has also been pointed out, [3-4], that methods that offer N-best recognition hypotheses can correct only a small percentage of the errors and it is difficult to display these N-best recognition hypotheses on the small screens of mobile devices. Speech-based correction methods that allow user to re-speak

X. Jiang, M. Ma and M. Rohs (Eds): Proceedings of WMMP 2010.

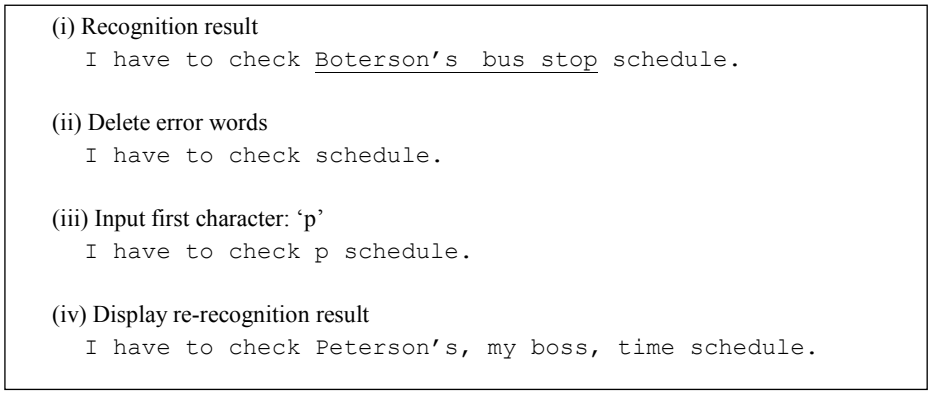


Fig. 1. An example of error correction

increase the burden on the user. Dynamic programming algorithm-based methods [5] calculate the similarity between a recognition result (a word) and words in a dictionary. The word that is most similar to the recognition result is selected as the correction result. This method is very simple, however its performance is low, especially when the error consists of multiple words. These methods have not been widely accepted for mobile use because of poor usability and performance. The most common method of correcting errors is the key input method; the necessary operations for the user are: 1. move cursor to the beginning of the error spot, 2. delete the error and 3. input the correct word or words. Obviously the number of key inputs increases with the number of errors and practicality is very low given the tiny keypads and small screens of mobile devices.

We have recently reported a re-recognition-based error correction method [6][7] that demands user indication of an error. The part of speech feature corresponding to the error is extracted and re-recognized by models in the local terminal. This method greatly improves usability because it reduces the number of key inputs and achieves more than 70% correction rate.

However, a certain number of error words remain, so the user still has to input key strokes to correct these errors. To reduce these input operations, a method that incrementally corrects the errors during the user's input is desired.

This paper proposes a novel completion method that reduces the key strokes needed to produce the correct text. Each character input by the user triggers re-recognition of the text that begins with these characters; each character is used to constrain the re-recognition process. This method incrementally corrects/refreshes the recognition result until the complete error is corrected.

The concept of auto word completion is widely used in text-based interfaces [8]. For example, mobile terminals provide a predictive input method that provides word candidates after the user inputs the first few characters. Compared to these text-based predictive methods, our proposed method benefits from two aspects: 1. it guarantees that the acoustical characteristics of corrected text match those of the speech input, 2, it can correct errors consisting of multiple words.

Figure 1 shows an example of the correction method and its result. The user spoke “I have to check Peterson's, my boss, time schedule”. Because the server has no user-specific knowledge, it fails to catch the name ”Peterson” and outputs an error around ”Peterson” (i). The user deletes the error words (ii), while the mobile terminal detects the error spot and extracts the corresponding speech feature, the user inputs the first character, “p” (iii), and at this time the mobile terminal corrects the error by reference to the user's dictionary (iv). This result shows that the proposed method offers greatly improved user usability.

The proposed method has the following advantages: 1. User inputs only as many characters as needed to correct all characters so that practicality is very high. 2. The corrected result is displayed and renewed whenever the user inputs a character, so the response is very fast. In addition, as re-recognition is performed only on the error part, the process load is light enough to be implemented on the mobile device, which has limited computational power. Furthermore, the indicated error is used to automatically extract the part of speech feature corresponding to the error. Therefore, the proposed method eliminates the need for re-speaking, which reduces the user's burden.

The remainder of the paper is organized as follows. We explain our error correction technique in the next section and then report some experiments on speech recognition. The paper concludes with a general discussion and issues related to future research.

2. Error correction for DSR with Search Space Constraint

2.1 System Description

The standard DSR system consists of a client and a server. The client front-end extracts features from the speech signal. The compressed features are then sent to the ASR (Automatic Speech Recognition) server. The ASR server performs speech recognition using these features and the recognition results are sent back to the client. Figure 2 shows the block diagram of the proposed error correction method. After the DSR process, also depicted in this figure, error correction will be performed on the received recognition result if needed. We apply context-based error correction [6] that utilizes the words that precede and follow the error spot. Basic context-based error correction consists of the following procedures.

1) Time alignment between the received text and the stored feature is performed and a time label indicating the start and end time obtained for each word. Because the text is given by the server, calculation of time alignment in the client device is very fast and accurate. This process can be omitted if the time label or similar information is included in the recognition result.

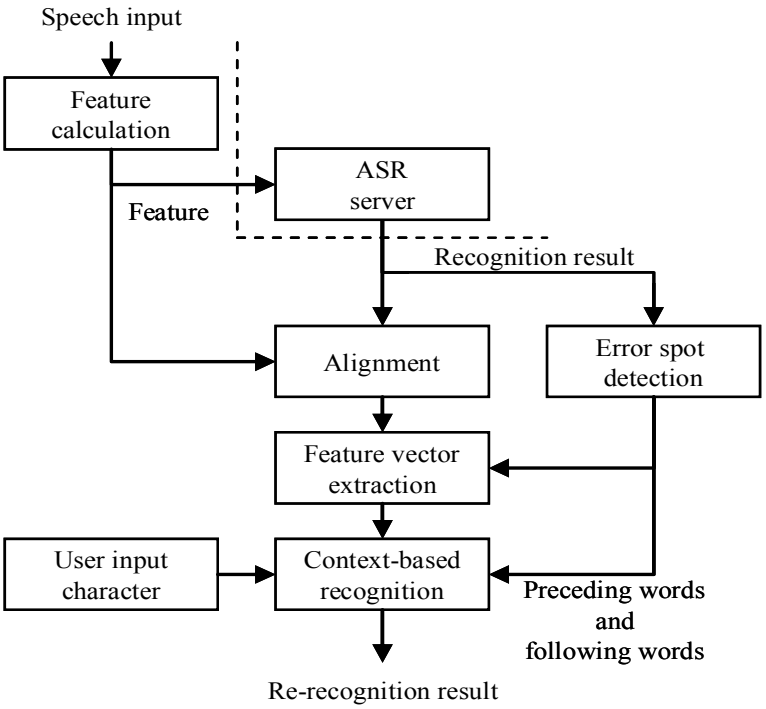


Fig. 2. Block diagram of proposed error correction

2) Detection of the start and end point of an error and identifying the words that precede and follow the error.

3) Extraction of the feature vector of the error spot from the stored feature. The vector includes the features corresponding to the preceding and following words as shown in Figure 3.

4) Recognition is performed on the extracted feature vector. The preceding words and following words are used to constrain the search space for re-recognition.

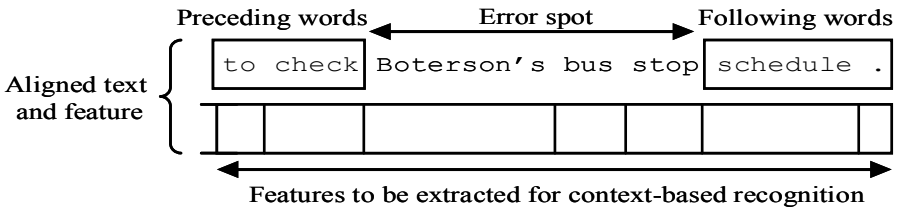
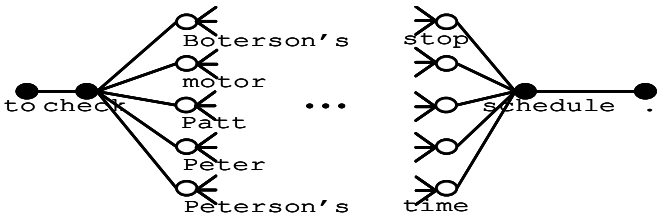
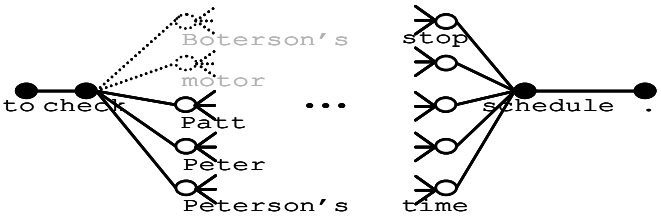


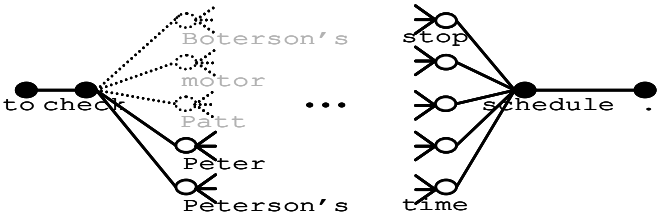
Fig. 3. Text and feature to be used for context-based recognition



(a) Search space constrained by preceding words and following words



(b) Search space further constrained by the character 'p'.



(c) Search space further constrained by the character 'e'.

Fig. 4. Search space limited by user input characters.

If the result is still wrong, the user's next key stroke is used for further search space constraint and re-recognition is performed again with the latest search space. This process is continued character-by-character until the correct recognition result is obtained or the entire error is corrected by the user input. Search space constraint is explained in the next subsection.

2.2 Search Space Constraint by User Input

Figure 4 shows examples of the search paths constrained by the user-input characters. The example shown in Figure 1 is used. Each node corresponds to a candidate word and the selected nodes are marked by black. The bold lines indicate the active paths. It should be noted that this figure depicts just the language domain for clarity but the acoustic search space is also constrained.

Figure 4(a) shows the basic correction method in which the search space is limited in range by the preceding and following words. Figure 4(b) shows the search space when user inputs th**e** character 'p' and Figure 4(c) shows the search space when the user inputs the second character 'e'. This method narrows the search space whenever the user inputs a new character so it is easy to find the correct text.

3. Experiment

3.1 Task and Data

To confirm the performance of the proposed method, we evaluated it by using 100 speech samples made by 12 speakers, 6 males and 6 females. The content of the test speech was taken from a News corpus and contained 1504 words. The average spoken utterance length was 5.7 seconds. We simulated a user-adapted language model trained by News data from the Mainichi newspaper corpus as it's content is close to that of the test data. This language model [9] contained 20k words and was assumed to be held by the mobile terminal. We simulated a server language model that was trained by data collected from WEB. The WEB language model [9] consists of a large corpus (60K) and is assumed to be installed on the server. N-gram language models are constructed based on the lexicon. Specifically, word 3-gram models are trained using back-off smoothing. The Witten-Bell discounting method is used to compute back-off coefficients. We implemented our method on the recognition engine Julius [9].

The speech signals were converted into a 25-dimension acoustic vector consisting of 12-dimension cepstral-mean-normalized MFCCs and their first derivatives, as well as normalized log energy coefficients [10]. The HMM used in our experiments was a 5 state, left-to-right HMM. The acoustical model was a speaker independent model trained by young or middle-aged adult speakers (150 sentences ×260 subjects).

3.2 Result of Server and Basic Error Correction

The baseline value of server recognition accuracy was 82.4%. There were 56 errors containing 264 erroneous words. An experiment was first performed wherein the subject specified the start and end points of each error and the result was 95.2 %. The remaining errors were used for evaluation. The causes of these error words include OOV (out of vocabulary) words and homonyms. These errors can be corrected by other well-known techniques [11][12]. We exclude these errors and use the remaining data for evaluating our correction method. The evaluation data consisted of 21 error spots with 26 words (some error spots consisted of multi-words). The error spots contained a total of 128 characters.

3.3 Result of Proposed Method

Figure 5 shows the performance of the proposed method. The x-axis is the error spot index. The y-axis plots R,T for each error spot. T is the total number of characters in the error spot and R is the input key number needed to produce the right text. For example in Figure1, T is 20 (the number of characters in "Peterson's, my boss, time"). If the user inputs the first two characters `p e`, and the error is corrected then R=2. We evaluated performance by two metrics: key reduction rate (KRR) and word correction rate (WCR). Key reduction rate is defined as KRR = 100%*R/T. WCR is the ratio of corrected words to total words. Corrected words means the correct text realized before the actual key strokes equaled the total number of erroneous characters (i.e. R<T).

The proposed method achieves 69.5 % KRR and 80.7 % WCR. It greatly reduces the number of user strokes needed for correcting the error words. It also can be seen that in most cases R is 1, which means that the user input just one character and all other characters were corrected automatically as in the example in Figure 1(iii) and (iv).

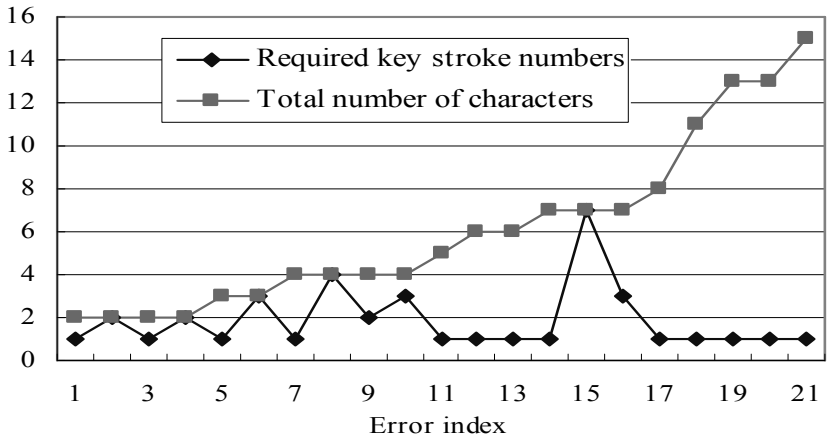


Fig. 5. Required key stroke numbers by the proposed correction method

4. Conclusions

We proposed a highly practical error correction method for mobile phones in the framework of DSR. The method reduces the number of key inputs needed to produce

the correct text by incremental correction in response to characters input by the user. Each character further constrains the feature space used by the re-recognition-based correction method to produce text that begins with these input characters. This method incrementally produces and displays the result as the user inputs each character. We confirmed that the proposed method reduces the keystroke number required for correction by 69.5% and errors by 80.7%.

Future research includes an evaluation on actual mobile devices, the effectiveness of user adaptation of the acoustical model, and tests on a larger database.

References

1. Karray, L., Jelloun, A. B., and Mokbel,C. :Solutions for Robust Recognition over the GSM Cellular Network , In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp. 261–264, (1998)
2. Suhm ,B., Myers ,B. and Waibel,A. :Multimodal Error Correction for Speech User Interfaces , In: ACM Transactions on Computer-Human Interaction, vol.8 no.1, pp.60-98, (2001)
3. Karat, C-M., Halverson ,C., Horn ,D. and Karat ,J.: Patterns of Entry and Correction in Large Vocabulary Continuous Speech Recognition Systems. In: Proc. CHI, pp.568-575 (1999)
4. Ogata and M. Goto : Speech Repair: Quick Error Correction Just by Using Selection Operation for Speech Input Interfaces, In: Proc. Interspeech, pp.133-136 (2005)
5. Allison. L.: Dynamic-programming can be Eager. Information Processing Letters, 43-pp.207 - 212 (1992)
6. Zhang. Z., Nakashima ,Y. and Naka ,N.: Error Correction with High Practicality for Mobile Phone Speech Recognition, In: Proceedings of Workshop on Speech in Mobile and Pervasive Environments. (2008)
7. Zhang, Z., Nakashima ,Y. and Naka ,N.: Error Correction via One Key Operation for Mobile Phone Speech Recognition, In: Proceedings of WMMP (2008)
8. Komatsu, H., Takabayashi,S., and Masui, T.: Corpus-based Predictive Text Input. In: Proceedings of the Third International Conference on Active Media Technology (2005)
9. Kawahara, T, Lee, A, Takeda, K, Itou, K, and Shikano, K. :Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository. In: Proc. International Conference on Spoken Language Processing, pp. 688--691, (2004)
10. Kawahara, T., Lee ,A., Kobayashi ,T., Takeda ,K., Minematsu ,N., Sagayama ,S., Itou ,K., Ito ,A., Yamamoto ,M., Yamada ,A., Utsuro ,T., and Shikano ,K. :Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition. In Proceedings of International Conference on Spoken Language Processing, pp. 476--479 (2000)
11. Honma ,S., Kobayashi ,A., Onoe , Sato ,K., S., Imai ,T. and Takagi ,T. :Speech Recognition with Out-of-Vocabulary Word Processing Using a Variable-Length Sub-Word HMM, IEICE technical report, vol.106, pp. 49-54, (2006)
12. Golding ,A., and Schabes ,Y. :Combining Trigram-based and Feature-based Methods for Context-sensitive Spelling Correction, In: Proceedings of Meeting of Association for Computational Linguistics, pp. 71-78, (1996)

Workshop Chairs:
Xiaoyi Jiang, Germany
Matthew Ma, USA
Michael Rohs, Germany

Program Committee:
Suchendra M. Bhandarkar, USA
Susanne Boll, Germany
Yung-Fu Chen, Taiwan
David Doermann, USA
Hamed Ketabdar, Germany
Christian Kray, UK
Jakob Eg Larsen, Denmark
Xiaobo Li, Sweden
Yuehu Liu, China
Michael O'Mahony, Ireland
Lucas Paletta, Austria
Marius Preda, France
Encrico Rukzio, UK
Andreas E. Savakis, USA
Shiva Sundaram, Germany
Rahul Swaminathan, Germany
Tan-Hsu Tan, Taiwan
Zheng-Hua Tan, Denmark
Steffen Wachenfeld, USA
Daniel Wagner, Austria
Hong Yan, Hong Kong, China
Zhipeng Zhang, Japan

Workshop Website:
http://cvpr.uni-muenster.de/WMMP2010/

Conference Website:
http://www.icpr2010.org

Inquiries:
Prof. Xiaoyi Jiang
Department of Mathematics and Computer Science
University of Münster
Germany
Email:
xjiang@math.uni-muenster.de

Important Dates:
Paper submission: 2010/4/1
Author notification: 2010/5/1
Early registration: 2010/5/14

CALL FOR PAPERS

The Second International Workshop on Mobile Multimedia Processing (WMMP 2010)

In Conjunction with
The 20th International Conference on Pattern Recognition 2010

August 22nd, 2010, Istanbul, Turkey

Scope:
This workshop aims at timely addressing the challenges in applying advanced pattern recognition, signal processing, computer vision and multimedia techniques to mobile systems, given the proliferating market of mobile and portable devices that have been widely spreading in both consumer (e.g. smartphones such as iPhone, music, mobile TV, digital cameras, HDTV) and industrial markets (e.g. control, medical, defense etc.). The proposed scope of this workshop includes, but not limited to, the following areas:

- Mobile speech, image and video processing
- Surveillance, biometric, authentication and security technologies in mobile environment
- Mobile visual search.,image retrieval and video streaming
- Multimodal pedestrian navigation systems
- Multimedia applications for automotive systems
- Mobile navigation, content retrieval, authentication
- Pervasive computing / context aware methodology and application
- Multimodal interfaces and visualization for mobile devices
- Handheld augmented reality
- Personalization and recommender systems in mobile environment
- Mobile oriented media processing for communication and networking
- Medical applications and bioinformatics in mobile environment
- Entertainment applications in mobile environment
- Mobile multimedia applications in geospatial information systems

The intended audiences of this workshop are primarily researchers in traditional pattern recognition and media processing techniques wanting to extend their work in the mobile domain, and those researchers in mobile and cross related fields wanting to explore latest achievement in pattern recognition field to expand their work.

Registration and Publication:
Attendees can register for the Workshop only for 100 Euros without having to register at the ICPR conference. All accepted papers will be published in the Workshop proceedings on CD-ROM. In addition, a related Special Issue with selected papers will be published in International Journal on Pattern Recognition and Artificial Intelligence.