MULTIMEDIA COMMUNICATIONS TECHNICAL COMMITTEE
IEEE COMMUNICATIONS SOCIETY
http://mmc.committees.comsoc.org/

MMTC Communications – Review

IEEE COMMUNICATIONS SOCIETY

Vol. 8, No. 4, August 2017

TABLE OF CONTENTS

Message from the Review Board Directors

Welcome to the August 2017 issue of the IEEE ComSoc MMTC Communications – Review.

This issue comprises three reviews that cover multiple facets of multimedia communication research including fine-grained textual image classification and recognition, 3D motion tracking in depth videos, and tiling of interactive panoramic vidoes. These reviews are briefly introduced below.

The first paper, published in IEEE Transactions on Mutimedia and edited by Bruno Macchiavello, investigated how to improve fine-gained classification based on both visual and textual information.

The second paper is published in IEEE Transactions on Multimedia and edited by Dr. Carl James Debono. It proposed to use depth videos obtained from Kinect 2.0 to estimate the heart rate and rhythm.

The third paper, published in IEEE Transactions on Multimedia and edited by Roger Zimmermann, investigate the fundamental tradeoff between maximizing user's experience and minimizing bandwidth requirement of video transmissions.

All the authors, nominators, reviewers, editors, and others who contribute to the release of this issue deserve appreciation with thanks.

IEEE ComSoc MMTC Communications – Review Directors

Pradeep K. Atrey
State University of New York at Albany, USA
Email: patrey@albany.edu

Qing Yang
University of North Texas, USA
Email: qing.yang@unt.edu

Wei Wang
San Diego State University, USA
Email: wwang@mail.sdsu.edu

Jun Wu
Tongji University, China
Email: wujun@tongji.edu.cn

Combining Textual content with Image Information for Fine-Grained Classification

A short review for "Words Matter: Scene Text for Image Classification and Retrieval"

Edited by Bruno Macchiavello

Sezer Karaoglu, Ran Tao, Theo Gevers, and Arnold W. M. Smeulders, "Words Matter: Scene Text for Image Classification and Retrieval," IEEE Transactions on Multimedia, vol. 19, issue. 5, p. 1063-1076, May 2017.

Fine-grained classification and recognition is receiving increasing attention among computer vision researchers [1-3]. Fine-grained classification aims to correctly separate images that posses very close appearance into different classes. Even for human observers, fine-grained classification tasks can require expert and domain specific knowledge. The authors of this work propose a fine-grained classification system for images of business places. Automatic recognition and indexing of business places can be useful in several practical scenarios. For example, it can be used to improve online services as Google street view or Google Map.

The main contributions of this paper are the following. First, this paper shows that textual cues can be use as complementary information to visual cues for fine-grained classification. Text in natural images can add meaning to an object or scene. In the particular case of business places, text can specify which business places serve drinks, food or what kind of service is provided. Second, the authors propose an efficient unsupervised word extraction method. And third, they annotated approximately 10K images from a dataset, which provides ground truth for word detection. Moreover, the authors also stated that they experimentally showed that high recall in word detection is more important than high f-score, for the applications consider in this work.

In order to extract the text cues from the image word segmentation and recognition is require. The presented word segmentation method aims to detect word boxes with high recall. To this end, it is propose the use of a complementary set of algorithms. Two previously proposed methods are used: text saliency [4] and maximally stable extremal regions (MSER) [5]. The text saliency algorithm constructs a saliency map using the image background. It is assume that the background pixel are uniform colored and that they contrasts with text regions. The method selects initial background seeds and grows these seeds iteratively until all background pixels are covered. MSERs define an extremal region as a connected component of which image values remain stable within the boundaries. The character candidates are obtained from the joint results of the MSER and text saliency algorithms However, text regions can still be missed, since images are capture under uncontrolled illumination conditions. Text regions may be influence by different photometric changes such as shadows. To compensate for this issue, the authors propose to compute the character candidates using a variety of color spaces containing a range of invariant properties. While this reduce the number of false negatives, the final character candidates may consist of non-character regions. Hence, the authors used features such as aspect ratio, size, solidity and contrast in order to filter non-character regions. The use of this features for this purpose has been reported previously [6]. The next step is to compute word box proposal. The word box are estimated based on five constraints: distance between characters, the ratio of vertical displacement and horizontal offset, The height ratio of two consecutive characters and if the bottom of a character is below the center of a contiguous one. Once the word box are defined, word recognition is performed through a convolutional neural network (CNN). The network takes a word box proposal b as input and products for each word w a probability of the word being present in the box. As a result each word box proposal is represented by a n-dimensional feature, where n is the number of words in the vocabulary. The authors used a previous model [7] of 88,172 words in the vocabulary.

For visual cues, the authors propose the following. The use of bag of visual words representation

and the use of the CNN ImageNet-pre-trained GoogleLeNet [8]. The pre-trained GoogleLeNet is tune in different manners in order to obtain a total of 7 (seven) visual cues. To fuse the visual and textual cues, kernel fusion is employed. Specifically, two kernels are computed separately, one for visual and one for textual, and the two kernels are summed to derive the final kernel matrix.

The authors provided results for the word box proposal evaluation, in which they indicate that the presented method yields slightly higher recall than previous works. More importantly the authors presented results for fine-gained classifications using text only cues, visual only cues and a combination of text and visual cues. The results indicate that the best performance is obtained by a combination of text and visual cues. Moreover, it appears that word-level textual cues can be more effective than character-level cues. For simulations the Con-text dataset was used [9], however in order to study the influence of precision and recall for word detection in the context of fine-grained classification, the authors annotated text regions for around 10K images within the dataset.

In conclusion, in this work it was presented a methodology to combine both visual and textual information in order to improve fine-gained classification. It was shown that word-level cues combined with visual information can improve performance during classification compare to the use of visual only cues. Nevertheless, this work focus specifically on business places. As future work, it could be interesting to investigate other scenarios of fine-grained classification where text information can also be use to improve performance.

References:

[1] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur, "Dog breed classification using part localization," in Proc. Eur. Conf. Comput. Vis., 2012, pp. 172–185

[2] S. Maji, E. Rahtu, J. Kannala M. Blaschko, and A. Vedaldi, "Finegrained visual classification of aircraft," CoRR, 2013. [Online]. Available: http://arxiv.org/abs/1306.515

[3] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-Grained Recognition without Part Annotations," CVPR, 2015.

[4] S. Karaoglu, J. C. van Gemert, and T. Gevers, "Object reading: Text recognition for object recognition," in Proc. Eur. Conf. Comput. Vis. Workshops Demonstrations, 2012, pp. 456–465.

[5] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in Proc. Brit. Mach. Vis. Comput., 2002, pp. 36.1–36.10.

[6] B. Epshtein, E. Ofek, and Y.Wexler, "Detecting text in natural scenes with stroke width transform," in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog., Jun. 2010, pp. 2963–2970.

[7] M. Jaderberg, K. Simonyan,A.Vedaldi, andA. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," CoRR, 2014. [Online]. Available: http://arxiv.org/abs/1406.2227

[8] C. Szegedy et al., "Going deeper with convolutions," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., Jun. 2015, pp. 1–9.

[9] S. Karaoglu, J. C. van Gemert, and T. Gevers, "Con-text: Text detection using background connectivity for fine-grained object classification," in Proc. ACM Int. Conf. Multimedia, 2013, pp. 757–760.

Bruno Macchiavello is an associate professor at the Department of Computer Science of the University of Brasilia (UnB), Brazil. He received his B. Eng. degree in the Pontifical Catholic University of Peru in 2001, and the M. Sc. and D.Sc. degrees in electrical engineering from the University of Brasilia in 2004 and 2009, respectively. Prior to his current position he helped develop a database system for the Ministry of Transport and Communications in Peru. He is and Are Editor for the Elsevier Journal Signal Processing: Image Communications. His main research interests include video and image coding, image segmentation, distributed video and source coding, multi-view and 3D video processing.

Using Depth Video for Hearth Rate and Rhythm Estimation

A short review for "Estimating Heart Rate and Rhythm via 3D Motion Tracking in Depth Video"

Edited by Carl James Debono

C. Yang, G. Cheung, and V. Stankovic "Estimating Heart Rate and Rhythm via 3D Motion Tracking in Depth Video," IEEE Transactions on Multimedia, vol. 19, no. 7, pp. 1625 – 1636, July 2017.

The growth in aging population and healthcare costs are changing the healthcare system towards individual-centered systems. This shift implies that medical monitoring systems are supplied to patients for continuous monitoring, allowing for early disease detection and timely response [1]. Most of the current physiological monitoring systems demand that users wear sensors using straps or sticky electrodes. This can discourage their use as they impact on the quality of life of the user by creating discomfort. Therefore, a contact-free solution is desirable to increase adoption and use of such health monitoring systems.

Computer vision based health monitoring systems provide a contact-free solution, where image processing tools are used to extract information from the captured images [2, 3]. Depth-image-based systems can also be used as they provide information about the distance between the objects captured and the depth sensor. This depth information is available per pixel. Some previous work in literature has shown that depth can be used for physiological monitoring [4, 5]. However, current depth capturing sensors have typically low bit-depth and suffer from noise. This makes it difficult to accurately measure very small physiological movements in patients.

The authors of the original paper use the depth video from a Kinect 2.0 to estimate the heart rate and rhythm. Blood pumped from the heart causes slight oscillations in the head, of the order of 5mm, with each beat due to Newtonian mechanics. This movement can be tracked and used to estimate the heart rate [6]. Given the small oscillatory movement that needs to be detected, the limitations of the depth sensing systems outlined above provide a clear challenge.

To mitigate these limitations, the authors of the original paper first restore the depth images by applying a joint bit-depth enhancement/denoising method, then tracking the head region to determine to determine its motion vectors which are projected onto its principal component for analysis, and finally estimating the heart rate and the rhythm using Welch power spectrum analysis and peak detection, respectively.

The reported system is made up of a Microsoft Kinect 2.0 camera connected to a standalone computer and positioned between 0.5m to 2m in front of the subject. The depth video captured using this setup has a resolution of 512×424 at a rate of 30 frames/second. The depth values are quantized and noise corrupted due to the depth precision and additive noise. A maximum a posteriori formulation is used to recover the original depth values. Graph-signal smoothness prior is used to support this process. A noise model is also derived and used in the denoising algorithm. The region of interest (RoI) that contains the subject's head needs to be specified and tracked from one video frame to the next. A template-matching solution is applied and a square area centered in the identified hemisphere is taken as the RoI. The kernelized correlation filter (KCF) [7] is then utilized to track the RoI in different frames. A motion prior feedback loop is implemented to ensure consistency between the motion data in the restored depth video and the tracking information. Mapping from the image space to the real-world 3D space is then performed to find the 3D motion. Principal component analysis (PCA) is used to project the 3D motion along the principal component to perform one-dimensional analysis that includes trend removal, bandpass filtering and motion denoising. The heart rate is then found using Welch power spectrum analysis and the heart rhythm is determined using peak detection.

The results reported in the original paper were performed on thirteen volunteers and compared

to the measurements of a portable finger pulse oximeter. Measurements were done with subjects facing different directions and some subjects had a mask on their face. Furthermore, the results are compared to other schemes. The reported results indicate that good estimates are obtained compared to the ground truth.

Digital medical monitoring systems that are non-invasive are becoming more and more important to support the current healthcare system and reduce costs. Computer vision and depth information have shown to provide promising results. However, more work is still needed to make these solutions viable. Measuring the motion of a subject's head introduces errors as the subject is expected to remain still during the measurements and that there is nothing supporting the head. This implies that other data from other sensing devices needs to be fused to combine features and extract more accurate data. Also, improved depth sensors that can provide superior quality depth video and more accurate algorithms that can improve the tracking accuracy of the head are needed.

References:

[1] X.-F. Teng, Y.-T. Zhang, C.C.Y. Poon, and P. Bonato, "Wearable medical systems for p-health," *IEEE Reviews in Biomedical Engineering*, vol. 1, pp. 62–74, December, 2008.

[2] F. Erden, S. Velipasalar, A. Z. Alkar, and A. E. Cetin, "Sensors in assisted living: A survey of signal and image processing methods," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 36–44, March 2016.

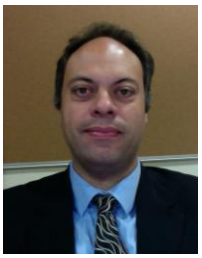
[3] G. Acampora, D. J. Cook, P. Rashidi, and A. V. Vasilakos, "A survey on ambient intelligence in healthcare," *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2470–2494, December 2013.

[4] C. Yang, G. Cheung, K. Chan, and V. Stankovic, "Sleep monitoring via depth video recording and analysis," in *Proceedings of the IEEE International Workshop on Hot Topics 3D*, July 2014, pp. 1–6.

[5] C. Yang, G. Cheung, V. Stankovic, K. Chan, and N. Ono, "Sleep apnea detection via depth video audio feature learning," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 822–835, April 2017.

[6] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3430–3437.

[7] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 583–596, March 2015.



Carl James Debono (S'97, M'01, SM'07) received his B.Eng. (Hons.) degree in Electrical Engineering from the University of Malta, Malta, in 1997 and the Ph.D. degree in Electronics and Computer Engineering from the University of Pavia, Italy, in 2000.

Between 1997 and 2001 he was employed as a Research Engineer in the area of Integrated Circuit Design with the Department of Microelectronics at the University of Malta. In 2000 he was also engaged as a Research Associate with Texas A&M University, Texas, USA. In 2001 he was appointed Lecturer with the Department of Communications and Computer Engineering at the University of Malta and is now an Associate Professor. He is currently the Head of the Department of Communications and Computer Engineering at the University of Malta.

Prof. Debono is a senior member of the IEEE and served as chair of the IEEE Malta Section between 2007 and 2010. He was the IEEE Region 8 Vice-Chair of Technical Activities between 2013 and 2014. He has served on various technical program committees of international conferences and as a reviewer in journals and conferences. His research interests are in wireless systems design and applications, multi-view video coding, resilient multimedia transmission, and modeling of communication systems.

Different Approaches for Tiling of Interactive Panoramic Video

A review for "Tiling in Interactive Panoramic Video: Approaches and Evaluation"
(Edited by Roger Zimmermann)

Vamsidhar Reddy Gaddam, Michael Riegler, Ragnhild Eg, Carsten Griwodz, and P°al Halvorsen," Tiling in Interactive Panoramic Video: Approaches and Evaluation", IEEE Transactions on Multimedia, vol. 18, no.9, pp. 1819–1831, September 2016.

The authors investigate and evaluate the tradeoffs when streaming panoramic video, specifically they are interested in the scenario when high-resolution video frames are decomposed into tiles and tiles are possibly transmitted at different quality levels. The authors investigate their methods in the context of large-scale panoramic videos that are created from multiple cameras. This work is very timely since recently there has been an increased interest in 360° video from both industry and academia. Some of this is driven by technological advances in that affordable consumer cameras are now available in the market which can output high-quality frame resolutions, for example at 4k.

The authors investigate the fundamental tradeoff between maximizing the viewer experience and at the same time minimizing the video transmission bandwidth. Panoramic videos tend to require high frame resolutions since a user in general only watches a sub-region of each frame, i.e., the user actively watches only a narrow field-of-view. Hence a possible optimization is to only stream the user's current region of interest (ROI), or at least only providing high quality video in the user's ROI, while streaming low resolution (or even no video) in the areas outside the ROI. On the other hand, with panoramic videos a user is allowed to interact with the content by panning or zooming and hence the ROI frequently and dynamically changes.

The authors recognize and investigate the multitude of different options that a system designer has in balancing the viewer's experience versus the required transmission bandwidth. The authors consider a cylindrical panoramic video that is divided into 8x8 tiles. As the delivery protocol the authors use dynamic adaptive streaming over HTTP (DASH) [1] as it supports the tiling approach well and is also now popularly used in industry. The authors then introduce four methods that differ in how they obtain the high quality tiles of the ROI versus the low quality tiles of the background area. The binary approach is the simplest method and it retrieves high quality tiles that overlap the ROI and low quality tiles everywhere else. The rescaled method generates a low quality thumbnail video of the complete panorama. Similarly to binary it uses high quality tiles in the ROI while filling in the remaining pixels of a panorama frame at the user's side by upscaling the thumbnail video. The third method utilizes viewing movement prediction and enlarges the high quality area based on such the predictions. Finally, the forth technique is called pyramid and it selects the qualities intelligently with a gradually decreasing quality according to the distance from the virtual camera.

One of the core contributions of this paper is the evaluation and comparison of the four mentioned methods. The authors start with a discussion of why existing quality metrics are difficult to apply in this context. Specifically, existing methods usually assume that all of a frame has been encoded in a similar way. However, the ROI of a tiled panoramic video may at times consist of both high quality and low quality tiles. The authors then study whether existing image similarity metrics such as SSIM [2] or OpenVQ are applicable by also performing an extensive user study. Specifically, the Microworkers platform was used for the user study and data was collected from 200 different participants.

The authors perform a very thorough evaluation and discuss various aspects of the four different delivery methods such as how they compare in bandwidth usage and the quality that can be achieved. Various conclusions are valuable for the designer of an interactive, panoramic video system. Foremost, using an approach with both high and low quality tiles is overall beneficial and can lead to substantial bandwidth savings. On the other hand, achieving good viewing movement

prediction is challenging and if it is not accurate it will lead to a higher bandwidth usage compared to the other methods. The pyramidal methods can provide both decent bandwidth savings and acceptable quality most of the time.

Overall I believe this study is very interesting and relevant, especially because of the growing support of 360° video by industry. Both Facebook and YouTube now allow users to upload and stream 360° video, however the underlying methods are not very resource efficient at this point. Studies such as this one are helping to improve the techniques such that panoramic video streaming will also be usable by smaller content providers.

References:

[1] MPEG-DASH Specification. ISO/IEC 23009-1:2012 Information technology – Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats, April 2012

[2] Z.Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004..

Roger Zimmermann is an associate professor with the Department of Computer Science at the School of Computing with the National University of Singapore (NUS) where he is also an investigator with the Smart Systems Institute (SSI). His research interests are in both spatio-temporal and multimedia information management, for example distributed systems, spatio-temporal multimedia, streaming media architectures, georeferenced video management, mobile location-based services and geographic information systems (GIS). He has co-authored a book, seven patents and more than two hundred conference publications, journal articles and book chapters in the areas of multimedia and spatio-temporal databases. He has received the best paper award at the ACM IWGS 2016 workshop and the IEEE ISM 2012 conference. He has been involved in the organization of conferences in various positions, for example TPC co-chair of ACM TVX 2017. He is an investigator with the NUS Research Institute (NUSRI) in Suzhou, China. Roger Zimmermann is an Associate Editor of the ACM Transactions on Multimedia journal (TOMM, formerly TOMCCAP) and the Multimedia Tools and Applications (MTAP) journal. He is a Senior Member of the IEEE and a member of ACM. For more details, see http://www.comp.nus.

Paper Nomination Policy

Following the direction of MMTC, the Communications – Review platform aims at providing research exchange, which includes examining systems, applications, services and techniques where multiple media are used to deliver results. Multimedia includes, but is not restricted to, voice, video, image, music, data and executable code. The scope covers not only the underlying networking systems, but also visual, gesture, signal and other aspects of communication. Any HIGH QUALITY paper published in Communications Society journals/magazine, MMTC sponsored conferences, IEEE proceedings, or other distinguished journals/conferences within the last two years is eligible for nomination.

Nomination Procedure

Paper nominations have to be emailed to Review Board Directors: Pradeep K. Atrey (patrey@ albany.edu), Qing Yang (qing.yang@unt.edu), Wei Wang (wwang@mail.sdsu.edu), and Jun Wu (wujun@tongji.edu.cn). The nomination should include the complete reference of the paper, author information, a brief supporting statement (maximum one page) highlighting the contribution, the nominator information, and an electronic copy of the paper, when possible.

Review Process

Members of the IEEE MMTC Review Board will review each nominated paper. In order to avoid potential conflict of interest, guest editors external to the Board will review nominated papers co-authored by a Review Board member. The reviewers' names will be kept confidential. If two reviewers agree that the paper is of Review quality, a board editor will be assigned to complete the review (partially based on the nomination supporting document) for publication. The review result will be final (no multiple nomination of the same paper). Nominators external to the board will be acknowledged in the review.

Best Paper Award

Accepted papers in the Communications – Review are eligible for the Best Paper Award competition if they meet the election criteria (set by the MMTC Award Board). For more details, please refer to http://mmc.committees.comsoc. org/.

MMTC Communications – Review Editorial Board

Multimedia Communications Technical Committee Officers

MMTC examines systems, applications , services and techniques in which two or more media are used in the same session. These media include, but are not restricted to, voice, video, image, music, data, and executable code. The scope of the committee includes conversational, presentational, and transactional applications and the underlying networking systems to support them.