LOW-OVERHEAD CONTENT-ADAPTIVE SPATIAL SCALABILITY FOR SCALABLE VIDEO CODING

Chia-Wen Lin[1,*], *Chia-Ming Tsai*[2], *Po-Chun Chen*[1], *and Li-Wei Kang*[3]

[1]Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
[2]Department of Computer Science and Information Engineering
National Chung Cheng University, Chiayi, Taiwan
[3]Graduate School of Engineering Science and Technology-Doctoral Program, and
Department of Computer Science and Information Engineering
National Yunlin University of Science and Technology, Yunlin, Taiwan
[*]Email: cwlin@ee.nthu.edu.tw

ABSTRACT

To support spatial scalability, the scalable extension of H.264/AVC (SVC) uses video cropping or uniform scaling to downscale the original higher-resolution (HR) sequence to a lower resolution (LR) one. Both operations, however, will cause critical visual information loss in the retargeted frames. The content-adaptive spatial scalability SVC coders (CASS-SVC) use non-homogeneous scaling to avoid critical information loss, which, however, requires to send additional side information to signal the decoder, thereby degrading coding efficiency significantly. To address the problem, we propose a low-overhead CASS-SVC coder consisting of three main modules: a mosaic-guided video retargeter, a side-information coder, and a non-homogeneous inter-layer predictive coder. Our experimental results demonstrate that, compared to existing CASS-SVC coders, our method cannot only well preserve subjective quality of important content in the LR sequence, but also significantly improves the coding efficiency of HR sequence.

Index Terms—Video adaptation, video retargeting, spatial scalability, scalable video coding.

1. INTRODUCTION

Network environments usually involve heterogeneous devices with various display abilities and channel bandwidths. While streaming a video through networks, video content needs to be adapted to match the heterogeneity of networks and user devices. Salable video coding, e.g., SVC [1], is an important technology to support the video content adaptation [1], [2]. To support various display resolutions and aspect ratios, SVC supports video cropping or uniform scaling to downscale the original HR sequence to an LR one. Different resolution videos are coded by individual video encoders and interlayer prediction is then used to reduce the redundancies between different spatial layers [2]. However, the flexibility and performance of the spatial scalability in SVC is still rather limited for high-quality video retargeting, since both video cropping and uniform scaling used in SVC lead to critical visual information loss in the retargeted frames. Recently, several content-adaptive video retargeting methods have been proposed [3]–[12]. These methods mainly aim to retain as much human interested regions as possible by trimming unimportant spatio-temporal content in the resized video, which can help enhance the performance of current SVC.

According to [3], content-adaptive video retargeting methods can be classified into discrete [4]–[7] and continuous approaches [8]–[12]. Seam-carving based methods are among the most representative discrete approaches [4]. Based on an energy function, such methods repeatedly remove a spatio-temporal surface until reaching the desired video resolution. Moreover, warping based methods [8]–[12] resize each frame by finding the optimal warping function of each patch in a continuous domain.

To preserve the visual information in the reduced-resolution video in SVC, recently a few content-adaptive coding schemes have been proposed to integrate the content-aware video retargeting with H.264 [13], [14] or SVC coders [15]–[17]. For example, Décombas *et al*. [13] proposed to integrate seam carving with H.264 for semantic video coding. Wang *et al*. [16], [17] proposed a content-adaptive spatial scalability framework to extend SVC. A warping based video retargeting method [9] is utilized in their framework to adapt the original full-sizes video to content-aware LR video.

The non-homogeneous scaling to obtain reduced-resolution video in the above content-adaptive coding schemes [13]–[17], however, requires to send additional side information (e.g., seam positions or warping parameters) to signal the decoder for properly reconstructing the original resolution video in the interlayer prediction process. The additional side information needs to be efficiently compressed so as to reduce its impact on coding efficiency. However, all the video retargeting approaches used in these methods require to store and send side information frame by frame, which consumes a significant amount of coding bits, thereby leading to significant coding efficiency degradation.

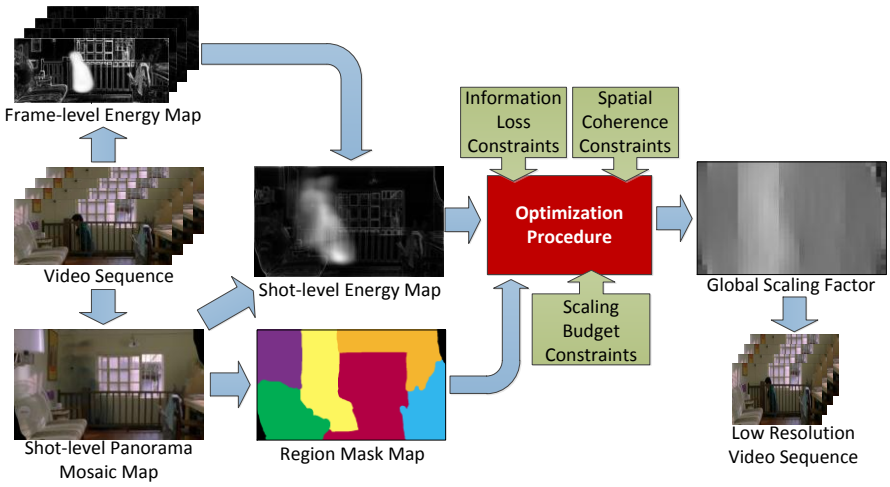


Fig. 1. Block diagram of the proposed mosaic-guided video retargeting method.

To address the problem, we propose a low-overhead content-adaptive SVC (CASS-SVC) coder which consists of three main modules: a mosaic-guided video retargeter, a side-information coder, and a non-homogeneous inter-layer predictive coder. Instead of sending per-frame side information used in existing methods [13]–[17], without sacrificing the visual quality in the retargeted video, our method only utilizes per-video-shot side information including shot-level global scaling maps and the spatial corresponding positions of individual frames to the panoramic mosaic. Since a video shot usually contains tens of hundreds of frames, the amount of side information is drastically reduced. The contribution of this paper is three-fold: (i) we propose a new low-overhead CASS-SVC coder; (ii) we propose new shot-based video retargeting schemes, preserving important visual content as well as maintain spatio-temporal coherence in retargeted video in low overhead; and (iii) we propose new non-homogeneous inter-layer prediction tools to achieve good coding efficiency.

2. SHOT-BASED VIDEO RETARGETING

We propose an efficient CASS-SVC coder which aims to preserve important visual information in the LR layer, as well as to reduce the overhead cost of sending side information for guiding the non-homogeneous inter-layer prediction at the decoder so that the coding efficiency of HR layer is not significantly degraded. Here, we modify our previous video retargeting method [11], [12] with some modification to downscale each video frame to its corresponding LR frame.

As shown in Fig. 1, by constructing shot-level panoramic mosaics, the shot-based video retargeting module non-homogeneously resizes each original frame to preserve important visual content in the frame while maintaining intra-frame spatial coherence and inter-frame temporal coherence with the help of the panoramic mosaics. More specifically, our method first performs shot boundary detection using the method proposed in [18] and then constructs a panoramic mosaic for each video shot. The panoramic mosaic image is then segmented into regions by using a semi-automatic segmentation tool [19]. Before deriving the scaling maps of individual frames, our method retargets the shot-level panoramic mosaic to the desired resizing ratio. Based on the shot-level panoramic mosaic, the region segmentation mask, a set of spatial coherence constraints, information loss constraints, and scaling budget constraints, we propose an iterative optimization scheme to obtain a shot-level global scaling map for the panoramic mosaic of a video shot, which is then used to derive the scaling maps of individual frames in the shot at both the encoder and decoder. More details of our retargeting scheme can be found in [11], [12].

Unlike our previous schemes [11], [12] which require to perform per-frame optimization to derive frame-level scaling maps, in this work, we embed the per-frame scaling budgets into the constraints for the iterative optimization of global scaling map. As a result, frame-level scaling maps can be directly computed from the global scaling map at both the encoder and decoder without resorting to the per-frame optimization. This does not only reduce the computation cost but also drastically reduces the amount of side information since only the global map along with a few alignment information need to be transmitted to the decoder.

3. SIDE INFORMATION CODERS

SVC exploits interlayer prediction and coding tools to enhance the coding efficiency for spatial scalability. The "I_BL" type macroblock (MB) is coded by a spatial scalability coding tool by which the HR block is reconstructed by adding prediction residues to the corresponding up-scaled LR block. Since LR frames in CASS-SVC are non-homogeneously downscaled from HR frames, additional side information is needed to signal the decoder to up-scale an LR frame correctly.

After obtaining the retargeted LR video, the global scaling maps and mosaic correspondence map are respectively coded by the proposed side information coder and sent to the decoder for guiding the reconstruction of the HR video. To further reduce the bitrate of the global scaling map, we impose an additional spatial constraint to limit the same column/row of the shot-level panoramic mosaic to share the same scaling factor. Since adjacent scaling factor values are close, DPCM (differential pulse-code modulation) is applied to remove the spatial redundancy among the values, followed by RLC (run-length coding) to encode the nonzero prediction residues. Finally, Huffman coding is used to remove the statistical redundancy among these 2D symbols generated by RLC.

Moreover, DPCM and Huffman coding are also used to encode the correspondence map of individual frames to the shot-level panoramic mosaic. Unlike the global scaling map coder, since the camera motion is unpredictable, RLC is not suitable to encode the correspondence values. After being coded by DPCM, the residues are directly encoded by using Huffman coding.

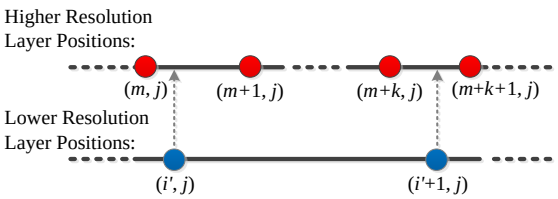


Fig. 2. Illustration of non-linear spatial mapping between two adjacent spatial layers in the horizontal direction.

4. NON-HOMOGENEOUS INTERLAYER PREDICTIVE CODERS

We modified the spatial scalability coding tools in SVC [1] to support the non-homogeneous inter-layer prediction used in our CASS-SVC.

4.1. The EL-BL Mapping Matrix

In our method, the EL-BL (enhancement layer-base layer) mapping matrix is designed to indicate the spatial correspondences of each frame from an HR frame to its retargeted LR frame. The correspondences from an HR frame to an LR frame is derived from the local scaling factors, $S_x^{(t)}$ and $S_y^{(t)}$, where $S_x^{(t)}(i,j)$ and $S_y^{(t)}(i,j)$ respectively denote the horizontal and vertical scaling factors for the (i,j)-th pixel in the t-th frame, derived by our video retargeting method described in Sec. 2.

Fig. 2 illustrates the nonlinear spatial mapping between two adjacent spatial layers in the horizontal direction. After horizontal warping, suppose the horizontal positions of the $(m+1,j)$-th to $(m+k,j)$-th pixels in the HR frame correspond to the fractional positions between (i',j) and $(i'+1,j)$ in the LR frame. Then, the corresponding EL-BL matrix values of the horizontal positions from $(m+1,j)$ to $(m+k,j)$ are all set to i'. The EL-BL matrix values of the t-th frame in x and y directions are computed by

$$\begin{cases} EL\_BL_x^{(t)}(m,n) = \text{Floor}\left(\sum_{i=1}^{m} S_x^{(t)}(i,n)\right) + 1, \forall n \\ EL\_BL_y^{(t)}(m,n) = \text{Floor}\left(\sum_{j=1}^{n} S_y^{(t)}(m,j)\right) + 1, \forall m \end{cases}, \quad (1)$$

where $\text{Floor}(\cdot)$ stands for the floor function which takes the largest integer smaller than the input.

4.2. Non-Homogeneous Interlayer Texture, Residue, and Motion Prediction

In the interlayer texture prediction in SVC, if an MB of the HR layer is coded as the I_BL MB type, the co-located blocks in the LR layer is up-scaled and subtracted from the corresponding MB of the HR layer. The prediction residue is then intra coded. In our method, the up-scaling operation is modified to support the non-homogeneous prediction as shown in Fig. 3(a).

Moreover, the interlayer residue prediction in SVC is performed when the corresponding spatial position in the LR layer is inter-coded. If the residue prediction is activated, the motion-compensated prediction residues of the co-located blocks in the LR layer is up-scaled and subtracted from the residues of the corresponding MB in the HR layer. The differences of residues are

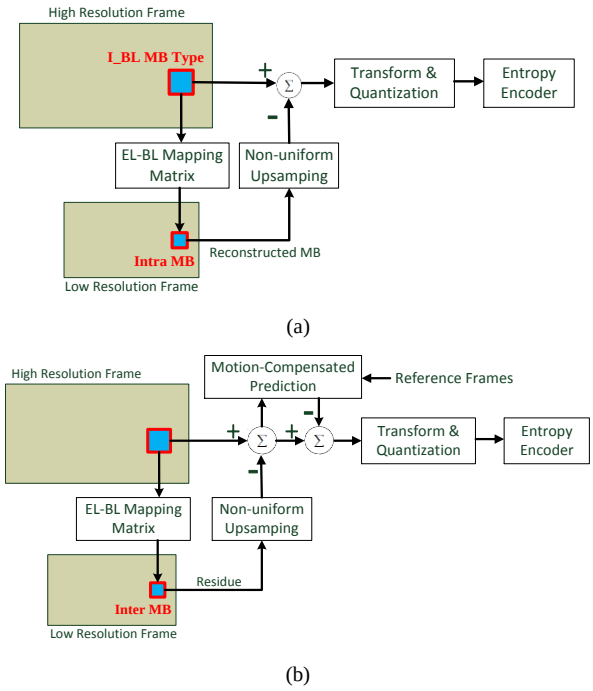


Fig. 3. Proposed non-homogeneous inter-layer predictive coders: (a) the texture prediction coder; and (b) the residue prediction coder.

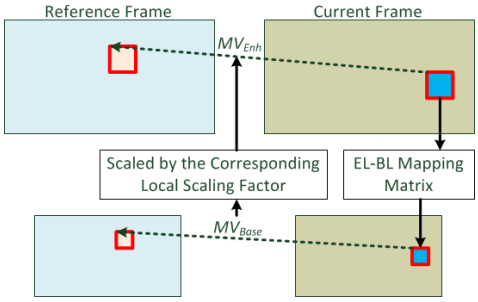


Fig. 4. Proposed inter-layer motion prediction coding structure.

then coded and embedded into the scalable bitstream. In our method, the upscaling operation is also modified to support the non-homogeneous prediction relations as shown in Fig. 3(b).

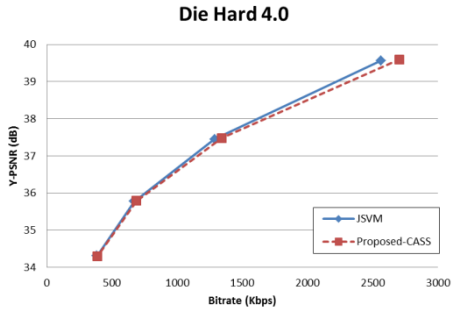
In addition, in SVC, when the motion prediction mode is activated, the motion vector of an MB in the HR layer is predicted either from the neighboring MBs of same layer or from the upscaled motion vector of the corresponding LR-layer MB. To obtain correct resampling ratios between two adjacent spatial layers for non-homogeneous upscaling, as shown in Fig. 4, our method uses the proposed EL-BL mapping matrix to retrieve the corresponding LR-layer motion vectors.

5. EXPERIMENTS AND DISCUSSION

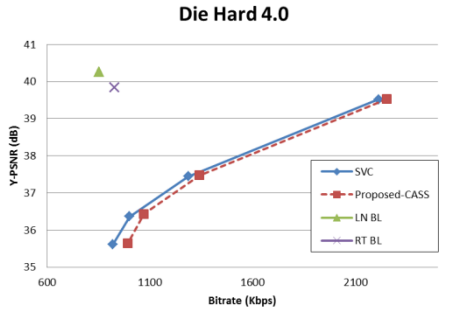
To evaluate the coding performance of the proposed CASS-SVC, we compare the rate-distortion (R-D) performance of our method with the SVC [1] and the method proposed in [16]. In the experimental settings, the width of each test sequence is halved. Each test sequence is coded using the hierarchical B-picture prediction structure with two spatial resolutions with a GOP (group of pictures) size of 16. Each HR sequence is downscaled by three

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

Fig. 5. Subjective quality comparisons of the downscaling results for (a) original video frame; by (b) uniform scaling used in SVC [1]; (c) retargeting scheme proposed by Krähenbühl *et al*. [9]; and (d) our video retargeting scheme.



(a)



(b)

Fig. 6. R-D performance (for the "Die Hard 4.0" sequence) comparisons between the proposed scheme and SVC [1] with the QP set: (a) (QP_{BL}^1, QP_{EL}^1); and (b) (QP_{BL}^2, QP_{EL}^2), where in the latter QP set, the QP value for the base layer is kept the same, and the R-D performances of the base layers of SVC (LNBL) and the proposed scheme (RTBL) are also plotted.

spatial downscaling schemes to obtain its LR sequences, including uniform downscaling for SVC [1], the retargeting method proposed in [9], used in [16], and our shot-based retargeting method for the proposed CASS-SVC, as exemplified in Fig. 5. The result demonstrates that our retargeting method can well preserve subjective quality of important content of an image and outperform the existing methods [1], [9] (used in [16]). More subjective evaluation results are available in [20].

To verify the performance of the proposed non-homogeneous inter-layer predictive coder under different quantization parameter (QP) settings, we designed two sets of QP settings: (QP_{BL}^1,QP_{EL}^1) = {(32,36), (28,32), (24,28), (20,24)} and (QP_{BL}^2, QP_{EL}^2) = {(24,36), (24,32), (24,28), (24,24)}. QP_{BL}^p and QP_{EL}^p denote the QP values for the LR (base) and HR (enhancement) layers in the p-th set, respectively. Figs. 5(a) and 5(b) compare the R-D performances of the HR-layer video between our method and SVC [1] using the two QP settings, respectively. Fig. 6(a) demonstrates that, no matter how the change of the QP value in the base layer, our method only leads to slight quality degradation, which is due to the additional overhead required for sending the side information to the decoder. As shown in Fig. 6(b), while increasing the coding bit-rate, the quality loss and effect of the overhead become insignificant even if

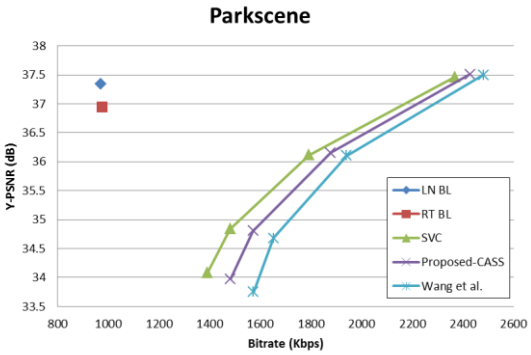


Fig. 7. R-D performance (for the "Parkscene" sequence) comparisons of the EL among the SVC [1], the proposed method, and the method in [16] (Wang et al.), where the QP set is set to (QP_{BL}^3,QP_{EL}^3)={(30,42), (30,38), (30,34), (30,30)} and the R-D performances of the base layers of SVC (LNBL) and the proposed scheme (RTBL) are also plotted.

the coding efficiency (instead of subjective visual quality) of base layer obtained by our method is worse than that of SVC.

We also compare the coding performance of the proposed coder with that of the coder proposed in [16] based on the same settings as in [16]. Fig. 7 compares the R-D performances of SVC [1], the proposed method, and the method in [16], evidently showing that the proposed method outperforms the method in [16], especially at low bitrates where the overhead cost becomes relatively significant. The main reason is the proposed method directly derives the frame-level scaling maps in a shot from the shot-level panoramic mosaic instead of sending per-frame scaling maps as in [16], resulting in much fewer side information.

6. CONCLUSION

We proposed a novel content-adaptive spatial scalability coding framework for SVC, which consists of three modules to preserve the important content in the retargeted LR video as well as to improve the coding efficiency for the HR layer. We have proposed a side information coder and efficient non-homogeneous interlayer prediction coding tools to achieve good coding efficiency in the HR layer. Thanks to our shot-based retargeting approach, our experimental results show that, compared to existing schemes, while maintaining comparable visual quality for the LR video, the proposed method consumes significantly lower overhead bitrate, thereby achieving better R-D performance.

7. REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable extension of the H.264/MPEG-4 AVC video coding standard," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sept. 2007.

[2] C. A. Segall and G. J. Sullivan, “Spatial scalability within the H.264/AVC scalable video coding extension,” *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 17, no. 9, pp. 1121–1135, Sept. 2007.

[3] A. Shamir and O. Sorkine, “Visual media retargeting,” in *ACM SIGGRAPH ASIA Courses* (*SIGGRAPH ASIA '09*), 2009, pp. 1–13.

[4] M. Rubinstein, A. Shamir, and S. Avidan, “Improved seam carving for video retargeting,” *ACM Trans. Graphics*, vol. 27, no. 3, pp. 16, 2008.

[5] C.-K. Chiang, S.-F. Wang, Y.-L. Chen, and S.-H. Lai, “Fast JND-based video carving with GPU acceleration for real-time video retargeting,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 11, pp. 1588–1597, Nov. 2009.

[6] D. Han, X. Wu, and M. Sonka, “Optimal multiple surfaces searching for video/image resizing - a graph-theoretic approach,” in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1026–1033, 2009, Kyoto, Japan.

[7] S. Kopf, J. Kiess, H. Lemelson, and W. Effelsberg, “FSCAV-fast seam carving for size adaptation of videos,” in *Proc. ACM Int. Conf. Multimedia*, pp. 321–330, Oct. 2009, Beijing, China.

[8] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video-retargeting," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2007, pp. 1–6, Rio de Janeiro, Brazil.

[9] P. Krähenbühl, M. Lang, A. Hornung, and M. Gross, “A system for retargeting of streaming video,” *ACM Trans. Graphics*, vol. 28, no. 5, 2009.

[10] Y.-S. Wang, J.-H. Hsiao, O. Sorkine, and T.-Y. Lee, “Scalable and coherent video resizing with per-frame optimization,” *ACM Trans. Graphics*, vol. 30, no. 4, 2011.

[11] T.-C. Yen, C.-M. Tsai, and C.-W. Lin, “Maintaining temporal coherence in video retargeting using mosaic-guided scaling,” *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2339–2351, Aug. 2011.

[12] C.-M. Tsai, T.-C. Yen, and C.-W. Lin, “Mosaic-guided video retargeting for video adaptation,” in *Proc. Conf. Applications of Digital Image Process. XXXIV, SPIE Optics+Photonics 2011*, Aug. 2011, San Diego, CA, USA.

[13] M. Décombas, F. Capman, E. Renan, F. Dufaux, and B. Pesquet-Popescu, "Seam carving for semantic video coding," in *Proc. Conf. Appl. Digital Image Process. XXXIV, SPIE Optics+Photonics 2011*, Aug. 2011, San Diego, CA, USA.

[14] M. Décombas, F. Dufaux, E. Renan, B. Pesquet-Popescu, and F. Capman, "Improved seam carving for semantic video coding," in *Proc. IEEE Workshop Multimedia Signal Process.*, pp. 53–58, Sept. 2012, Banff, Canada.

[15] T. M. Bae, T. C. Thang, D. Y. Kim, Y. M. Ro, J. W. Kang, and J. G. Kim, "Multiple region-of-interest support in scalable video coding," *ETRI J.*, vol. 28, no. 2, pp. 239–242, Apr. 2006.

[16] Y. Wang, N. Stefanoski, M. Lang, A. Hornung, A. Smolic, and M. Gross, “Extending SVC by content-adaptive spatial scalability,” in *Proc. IEEE Int. Conf. Image Process.*, Sept. 2011, pp. 3493–3496, Brussels, Belgium.

[17] A. Smolic, Y. Wang, N. Stefanoski, M. Lang, A. Hornung, and M. H. Gross, “Non-linear warping and warp coding for content-adaptive prediction in advanced video coding applications,” in *Proc. IEEE Int. Conf. Image Process.*, Sept. 2010, pp.4225–4228, Hong Kong, China.

[18] C.-W. Su, H.-Y. M. Liao, H.-R. Tyan, C.-W. Lin, D.-Y. Chen, and K.-C. Fan, “Motion flow-based video retrieval,” *IEEE Trans. Multimedia*, vol. 9, no. 6, pp. 1193−1201, Oct. 2007.

[19] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, Sept. 2004.

[20] NTHU Video Scaling project. [Online]. Available: http://www.ee.nthu.edu.tw/cwlin/cass_svc/cass_svc.htm.